NBER WORKING PAPER SERIES

WHEN IS DISCRIMINATION UNFAIR?

Peter J. Kuhn
Trevor T. Osaki

Working Paper 30236
http://www.nber.org/papers/w30236

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2022

The authors thank Catherine Weinberger and participants in the 2022 Trans-Pacific Labor Seminar and a seminar at Drexel university for helpful comments. This study and a pre-analysis plan were pre-registered in the AEA RCT Registry, under ID number AEARCTR-0006409. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

When Is Discrimination Unfair?
Peter J. Kuhn and Trevor T. Osaki
NBER Working Paper No. 30236
July 2022
JEL No. J71

ABSTRACT

Using a vignette-based survey experiment on Amazon's Mechanical Turk, we measure how people's assessments of the fairness of race-based hiring decisions vary with the motivation and circumstances surrounding the discriminatory act and the races of the parties involved. Regardless of their political leaning, our subjects do not distinguish between taste-based and statistical discrimination, but they react in very similar ways to other aspects of the act, such as the quality of information on which statistical discrimination is based. Compared to conservatives, moderates and liberals are much less accepting of discriminatory actions, and consider the discriminatee's race when making their fairness assessments. We describe four simple models of fairness –utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU)-- and show that the latter two are inconsistent with major patterns in our data. Instead, we argue that a two-group model in which conservatives care only about race-blind rules (RBRs), while moderates and liberals care about both RBRs and utilitarian ethics can account for the main patterns we see.

Peter J. Kuhn
Department of Economics
University of California, Santa Barbara
2127 North Hall
Santa Barbara, CA 93106
and IZA
and also NBER
pjkuhn@econ.ucsb.edu

Trevor T. Osaki
tosaki@ucsb.edu

1. Introduction

A large literature has studied the prevalence, magnitude and causes of discrimination based on characteristics that include race and gender. Another rapidly growing literature has studied the conditions under which people perceive income and pay inequality as fair or unfair, and has demonstrated that these fairness perceptions can have strong effects on peoples' economic and political behavior (Alesina and La Ferrara, 2005; Lefgren et al., 2016; Almas et al., 2020; Dube et al. 2021). Motivated by both these literatures, this paper studies whether and when people perceive discrimination as unfair, a question that has received very little attention among economists.

To answer these questions, we use a vignette-based survey experiment on Amazon's Mechanical Turk (MTurk) to measure people's assessments of the fairness of race-based hiring decisions. The vignettes illustrate canonical examples of statistical and taste-based discrimination, with both Black and White recipients of discrimination (*discriminatees*). In addition, the scenarios have varying levels of *justifiability*, i.e., varying motivations for the discriminatory act which we expect will make the actions more or less socially acceptable. The goals of our analysis are, first, to measure the effects of three types of factors on the perceived fairness of a discriminatory act in a broad sample of Americans: the characteristics of the respondent; the *motivation* for discrimination (e.g., tastes versus statistical); and the identity of the discriminatee (Black versus White). Second, we assess the ability of four different models of perceived fairness to account for the patterns we observe.

Our main findings are as follows. First, despite many economists' interests in distinguishing between taste-based and statistical discrimination, our subjects do not perceive a meaningful distinction between the fairness of these two types of discrimination: When faced with the same employer action (i.e., the employer rejected an equally qualified applicant because of that person's race), subjects do not care whether the decision was made because of racial animus or an expected productivity difference. Second, subjects *do* care about other aspects of employers' motivations. Specifically, taste-based discrimination by employers is seen as substantially less fair when it is based on the employer's own tastes than on the tastes of the employer's customers, and statistical discrimination is seen as less fair when it is based on low-quality information about relative group productivity, compared to higher-quality

information. Notably, the effects of these motivational factors on perceived fairness are widely shared: they are the same across all political groups, and they do not depend on the race of the discriminatee.

Third, respondents' assessments of the fairness of discriminatory actions differ dramatically by their political orientation. On average, conservatives rate the discriminatory actions depicted in our scenarios as slightly more fair than unfair, regardless of the race of the discriminatee. Moderates and liberals rate the discriminatory acts we describe as unfair; also, in contrast to conservatives, moderate and liberals exhibit a *discriminatee race effect*: they disapprove more of anti-Black than anti-White discrimination.

Fourth, among the four models of perceived fairness we consider –utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU)– the latter two are inconsistent with some major empirical patterns in our data. Specifically, racial in-group bias is inconsistent with the fact that, on average, respondents from all racial groups view anti-Black discrimination as less fair than anti-White discrimination. This effect is, in fact, especially strong and significant among White respondents, though we cannot reject that its magnitude is equal across all racial groups. We reject belief-based utilitarianism –a model in which, for example, conservatives' beliefs that Black people have equal or greater economic opportunities than White people can explain conservatives' assessment of anti-Black discrimination as fair-- in part because these beliefs (which are common) cannot account for the large political gap in fairness perceptions. The predictions of the BBU model are also at odds with how respondents of all political leanings rate the fairness of anti-White discrimination.

Turning to our two remaining fairness criteria –(simple) utilitarianism and race-blind rules (RBRs)-- we argue that a two-group model in which conservatives care only about RBRs, while moderates and liberals care about both RBRs and utilitarian ethics can account for most of the patterns we see. Finally, we use within-subject treatment variation in the race of the discriminatee to assess the relative weight moderates and liberals place on the two fairness criteria they appear to care about, finding that they place roughly equal weight on utilitarianism versus RBRs when forced to choose between them.

Our paper connects to a literature in labor and personnel economics that uses models of fairness to interpret the effects of pay inequality on effort, job performance and satisfaction, wage satisfaction, and

quits (Charness and Kuhn 2007; Abeler et al. 2010; Card et al. 2012; Charness et al. 2015; Bracha et al. 2015; Cohn et al. 2015; Breza et al. 2017; Cullen and Perez-Truglia 2018; Dube et al, 2019; Fehr et al. 2021). Some of these authors have argued, for example, that productivity-related wage differentials are seen as fairer than differentials attributed to other factors, such as luck (Abeler et al., 2010; Breza et al., 2017). We also connect to a related literature in experimental and personnel economics on the effects of the *intentions* behind an economic action on its perceived fairness (Charness and Levine 2000; Offerman 2002; Abeler et al. 2010; Breza et al. 2017). In a variety of contexts, including layoffs and within-firm pay inequality, these authors show that people's reactions to the same action vary dramatically with the reasons why the action was taken. None of these authors, however, consider the effects of the intentions behind a *discriminatory* act on its perceived fairness.[1]

A related literature in sociology has studied peoples' assessments of the fairness of income differentials, in many cases focusing on income gaps between women and men (Jasso and Rossi 1977; Auspurg, Hinz, and Sauer 2017; Jasso, Shelly and Webster 2019; Sauer 2020). Like us, these studies consider a number of implicit criteria people might use to judge income differentials; these criteria include *need* and *impartiality*, which roughly map into our utilitarian and RBR models. To our knowledge, however, this literature has not considered the perceived fairness of discriminatory actions.[2]

Our research also relates to some recent papers that study the effects of peoples' *beliefs* about the causes of inequality on their support for policies that redistribute income and opportunities, both overall (Alesina et al., 2020) and specifically on racial basis (Haaland and Roth 2021; Alesina et al. 2021). The latter two papers find that beliefs about the causes of racial inequality are highly correlated with support for both race-based policies like affirmative action; these beliefs also account for much of the partisan

[1] In fact, we are aware of only one other study that elicits peoples' assessments of the fairness of discriminatory acts: Feess et al. (2021) use vignettes similar to ours to compare subjects' views of anti-female versus anti-male discrimination. Barr, Lane, and Nosenzo (2018) use an allocator-game lab experiment to elicit second-order beliefs (which discriminatory acts do *others* see as fair?) of British university students. Our focus on first-order beliefs is motivated, in part, by the high level of political polarization in the United States. In such contexts --where social norms are contested—there could be large differences between first- and second-order perceptions of fairness, with the latter being highly sensitive to the identity of the persons whose beliefs the subjects are asked to predict.

[2] One recent sociology paper studies how peoples' willingness to engage in (hypothetical) acts of statistical discrimination can be manipulated. Tilcsika (2021) finds that exposing subjects with managerial experience to the theory of statistical discrimination increased the extent to which they relied on gender in a hiring simulation.

divide in policy support. Informational treatments designed to change people's beliefs, however, have limited effects on policy support.

Our paper differs from these papers in two main ways; the first is that we study a different outcome. Specifically, we focus on how our respondents assess the fairness of discriminatory *actions* taken by private individuals (employers in our case), not on respondents' expressions of support for various public policies. Second, and closely related, we consider a broader set of implicit fairness models that people might use to assess either actions or policies. Specifically, we will show that peoples' fairness assessments depend not only on an action's consequences (implicit in utilitarian assessments of public policies) but also on the actor's *intentions.* Intentions, and *rules* –i.e. a desire to apply a consistent of rules when mapping actions into fairness levels-- play important roles in non-consequentialist ethics such as those studied by Andreoni et al. (2019). A person's intent also plays a prominent role in criminal and civil law. In the paper we show that expanding the set of fairness models to include these considerations provides a more complete accounting of which types of discriminatory acts are perceived as fair or unfair.

Considering non-consequentialist factors may also provide a more complete accounting of which public policies are seen as fair. For example, a restrictive immigration policy might be seen as more fair if it was perceived to be motivated by a sincere desire to protect the earnings of low-income native workers than if it was motivated by racial animus. To our knowledge, economists have not yet studied the effects of policymakers' perceived motivations on how observers judge the fairness of their policies.

The rest of the paper proceeds as follows. Section 2 discusses the survey design, data collection, and sample characteristics. Section 3 presents some facts about fairness perceptions: How do perceptions vary with respondent characteristics, survey treatments, and the respondents' decision environment (i.e., the number and type of previous treatments encountered)? Section 4 describes four simple models of fairness and compares their implications to subjects' aggregate response patterns in Stage 1 of our survey.[3] Section 5 uses within-subject treatment variation between Stage 1 and 2 of the

[3] In Section 2.4 we show that question order effects matter for our *race* treatments, but not for our other two treatments (*motivation* and *justifiability*). Because the *race* treatments do not change within Stage 1 of the experiment, Sections 3 and 4 of the paper use only Stage 1 experimental data, in order to paint a clean picture of how subjects respond to all our treatments.

survey to study how subjects trade off their preferences for utilitarianism and race-blind rules when those objectives conflict. Section 6 concludes.

2. Survey Design

2.1 Overview

Before starting our survey, all our subjects are informed that they will be exposed to four scenarios, with the proviso that "Some of these scenarios may seem realistic to you; others may seem unrealistic." We also told subjects that only very limited information about each scenario will be provided. Nevertheless, subjects were asked to "please give us your reaction to [the scenarios] if they were to happen, based on the information that has been provided". The goal of these statements was to clarify that we want respondents to assess the *fairness* of the hypothetical interactions (and not their realism or their likelihood of occurring).[4]

Next, our respondents read and react to four vignettes that describe discrimination in a hiring decision. These vignettes randomly vary the race of the discriminator and discriminatee, as well as the *motivation* for the discriminatory action. The motivation could either reflect *taste-based* or *statistical* discrimination. Within each of these two broad categories are subcases in which discrimination is *more* versus *less* justifiable, where *justifiability* refers to how we hypothesized respondents would react. Specifically, *less*-justifiable taste-based discrimination is based on the employer's own distaste for people of a particular race, i.e. employer discrimination. More-justifiable taste-based discrimination by the employer occurs when the employer is accommodating their customers' distastes toward a particular race, i.e. customer discrimination. *Less*-justifiable statistical discrimination is based on low quality information – *hearsay* – about the relative productivity of workers from different racial groups.[5] Finally, statistical discrimination is categorized as *more* justifiable if it is based on high-quality, unbiased information about the relative productivity of different racial groups for this job.

[4] Complete instructions are provided in Appendix 1.
[5] The *less* and *more* justifiable variants of statistical discrimination correspond to what Bohren et al. (2019) describe as "inaccurate" and "accurate" discrimination, respectively. In Bohren et al.'s case, the accuracy refers to the decisionmakers' prior beliefs about a group's productivity level. For example, inaccurate beliefs could be based on stereotypes.

Each vignette describes an employer, randomly named "Michael" or "Andrew," who is making a hiring decision between a White and a Black applicant.[6] In all cases, the (unnamed) applicants are identical in all respects except their race. As mentioned, the employer's race can vary between vignettes, and he is always depicted as selecting the worker from his racial in-group. After each vignette, respondents are prompted to rate the fairness of the employer's hiring decision on a seven-point scale. Since we are interested in measuring respondents' first-order beliefs, none of these elicitation questions are incentivized.

The four scenarios encountered by each respondent are presented in two Stages. Each Stage is assigned a *race* and *motivation* treatment. The race treatments, Black and White, indicate the *race* of the discriminatee while *the motivation* treatments, Taste and Statistical, indicate the type of discrimination. Although respondents switch race and/or *motivation* treatments between Stages, respondents encounter both the *less*- and *more* justifiable subcase of either type of discrimination within each one. These *justifiability* treatments are administered in random order. Furthermore, the name of the employer also switches between the Stages.[7] Our survey concludes with Stage 3, which asks all subjects the same background questions.

2.2 Scenarios and Fairness Assessments

To illustrate how our fairness assessments work, we next describe Stage 1 of the survey for subjects who are assigned to the TB (Taste, Black) treatment combination. To introduce this Stage, we first tell subjects they will encounter two scenarios which share many common elements but contain some differences; we also say that the differences have been underscored to make them easier to pick out. The subjects then read and assess the *less* or *more* justifiable forms of the Taste discrimination scenario with a Black discriminatee in random order. The *less* justifiable form of taste discrimination is motivated by the employer's own tastes:

[6] Michael and Andrew appear to be among the most common male names that are relatively race-neutral. Between 2011 and 2016, they ranked in the top 2-6 names for White people and the top 6-12 names for Black people in New York City birth names. We use Michael in all our exhibits of survey instruments in the paper.
[7] Specifically, the name of the employer is randomized (50/50) to be either "Michael" or "Andrew" in both vignettes of Stage 1. In Stage 2, the name of the employer switches to the other, unused name for all respondents.

> Michael, who is White, is about to hire his first customer representative for his business after a few years of carrying that role alongside his managerial duties. He has interacted with a number of Black people during his education and work experience. While all of his interactions with Black people have been polite and professional, he just didn't enjoy interacting with them.
>
> For his new hire, Michael has to choose between two applicants whose resumes, interviews and references are all of equal quality, one of whom is Black and one who is White. Michael decides to hire the White worker in order to avoid interacting with a Black employee.

The *more* justifiable form is identical, except for the following underscored sections:

> He has conducted focus groups with a substantial share of the people who frequent his business. Many of these customers tell Michael that they do not like interacting with Black people and would be hesitant about continuing to support his business if he employed them. Michael himself is just as happy to interact with Black workers as with workers of other races.
>
> Michael decides to hire the White worker, in order to avoid losing sales to customers who do not want to interact with Black representatives.

After each scenario, the respondent is asked to "indicate the extent to which you thought that Michael's hiring decision was fair" on a seven-point scale, where 1 was "very unfair", 4 was "neither fair nor unfair", and 7 was "very fair".

In Stage 2, respondents encountered two more scenarios in which either the *race* of the discriminatee (Black or White), the *motivation* for the discrimination (Tastes versus Statistical) or both of these were changed. White scenarios were identical to Black scenarios except that the races of the discriminator (Michael) and the discriminatee are reversed. *Less* justifiable statistical discrimination was based on low-quality information (hearsay from a single, uninformed source) about relative group productivity, and *more* justifiable statistical discrimination was based on higher-quality information (quantitative information from substantial sample of other employers). To anticipate, we will find that the

circumstances summarized by these *justifiability* treatments have large effects on our subjects' fairness ratings, within both the Taste and Statistical types of discrimination.

The fact that these precise circumstances matter for fairness raises an important issue about how we can compare the perceived fairness of Taste versus Statistical forms of discrimination; essentially our comparisons hinge on the assumption that the average 'severity' or justifiability of our two Taste scenarios is the same as our two Statistical scenarios While we cannot think of any non-tautological way to test this assumption, we designed the more-justifiable versions of our Taste and Statistical discrimination scenarios to be the mildest clear examples of each type we could think of; we also made the difference between their more and less-justifiable variants equal in magnitude from our own perspective. Our intuition on the latter issue seems to be confirmed by the data, which show a very similar justifiability gap within the Taste versus Statistical treatments.

In addition to raising sample size – while preserving the option to use only each subject's first treatment for pure cross-subject comparisons – our motivation for exposing subjects to four scenarios has two additional benefits: First, it allows us to assess the relevance of race-blind ethics in a context where individual subjects are exposed to discriminatees of different races. Second – at least to the extent that the (randomly assigned) first *race* treatment subjects encounter affects their beliefs about the experimenters' political preferences — it allows us to test for experimenter demand effects.

2.3 Common Questions, Randomization, and Representativeness

After assessing the four scenarios in Stages 1 and 2, the subjects entered Stage 3 of the survey, where they answered a common set of questions. This included a question where subjects were asked to assess the relative "economic opportunities available to Black and White people" in the United States on a seven-point scale. We also collected information on the subjects' age, education, race, gender, and political affiliation.

Turning to our randomization approach, in Stage 1 of the fairness assessments, subjects were assigned with equal probability to one of the four possible treatment combinations: SW, TW, SB, and TB (where S, T, W and B are for statistical, taste, White and Black). In Stage 2, respondents were randomly assigned to one of the three combinations they did not encounter in Stage 1. Thus, as illustrated in

Appendix A2.1, two thirds of the respondents encountered a switch in the discriminatee's race, and two thirds encountered a switch between tastes-based and statistical discrimination. Within each of the two Stages, respondents encountered the *less and more* justifiable forms of discrimination in random order.

On September 21, 2020 we pre-registered our survey design and procedures, and posted a pre-analysis plan. Our survey was administered to a sample of MTurk workers between September 22 and October 6, 2020.[8] Subjects were given one hour to complete the survey and were informed that we expected the task to take about 15 minutes. Conditional on completing the entire survey, subjects were paid $5.[9] A few measures were taken to improve the accuracy and representativeness of the responses. First, respondents were required to have a U.S. address. Second, to further discourage foreign workers from participating, the survey was launched during U.S. Pacific daylight hours on weekdays. Third, MTurk workers were required to have a 95 percent approval rating to discourage robots (i.e., automated responses). Fourth, the survey included a CAPTCHA question to further discourage robots. Fifth, respondents were exposed to each vignette for at least 30 seconds before being allowed to submit their fairness assessment. Sixth, Appendices 7 and 8 re-weight all our main estimates to reflect the demographic mix in the American Community Survey and the political mix in the General Social Survey, with very similar results. Finally, Appendix P3.2 replicates some of our main results for a subset of 'thoughtful' respondents –those who took more than the median amount of time to complete the survey, again with very similar results. Some additional data cleaning resulted in a final count of 642 responses for the survey sample.[10]

Finally, we were concerned that the representativeness of our results would be affected by the civil unrest following the murder of George Floyd on May 25, 2020, which led to a mainstream conservation on systemic racism in the U.S.; it seems reasonable to ask whether these events may have primed our respondents to answer our questions in an unusual way. To check for this possibility Figure

[8] We re-weight our main results to match American Community Survey and General Social Survey data in Appendices 7 and 8. For additional discussions of the representativeness of MTurk samples, see Kuziemko et al. (2015) and Arechar et al. (2017).
[9] In comparison, the average effective hourly rate on MTurk is about $4.80 (Kuziemko et al., 2015). The average actual survey completion time for our subjects was 11.5 minutes.
[10] Additional details on these procedures, and summary statistics on the sample's characteristics (race, gender, education, political orientation, and location in the U.S.) can be found in Appendix 2.2.

A2.2.1 presents online search trends for related keywords, including *Black Lives Matter* and *racism* during the spring and summer of 2020. These trends show that searches for these terms had diminished dramatically by the time of our survey, suggesting that this type of priming may not be a significant issue for our respondents.

2.4 Question Order Effects

In Appendix 3, we demonstrate that question order effects are absent from our survey in two distinct senses. First, as shown in Appendix 3.1, there is no time trend in fairness assessments across the four scenarios encountered by each respondent —respondents become neither more nor less accepting of discrimination as they are asked additional questions about it.[11] Second, the order in which respondents encounter the Tastes versus Statistical and the *more v*ersus *less justified* scenarios does not affect their fairness ratings. In Appendices A3.2 and A3.3, this is illustrated three different ways: First, we show that subjects' subsequent assessments of a given type of discrimination (e.g., Taste) do not depend on which type (Tastes or Statistical) they encountered previously. Second, we cannot reject that the fairness ratings *changes* of subjects who switched from, say, a *more to a less justified* treatment were equal but opposite in sign to subjects who switched in the opposite direction. Finally, for both the type of discrimination and the *justifiability* treatments we show that within-subject, between-subject and pooled fairness regression estimates are statistically indistinguishable from each other.[12]

The one treatment that does, however, affect subjects' subsequent fairness assessments is the *race* of the discriminatee. As we document in Appendix A3.4, our respondents were considerably more tolerant of anti-Black discrimination in Stage 2 if they were randomly exposed to a White discriminatee in Stage 1 (compared to a Black discriminatee in Stage 1). In the next two Sections of the paper (3 and 4), we will eliminate the influence of order effects by relying only on data from Stage 1 of the survey: There is no within-subject variation in the *race* treatment during Stage 1 (or during Stage 2), because

[11] Recall that treatments are assigned in a balanced way across the four scenarios each respondent encounters, so aggregate comparisons of fairness ratings over time are not contaminated by changes in the mix of scenarios people encounter.

[12] Within-subject estimates regress fairness on a treatment indicator plus respondent fixed effects. Between-subject estimates are pure cross-section regressions using data from the first treatment each respondent encountered only. Pooled estimates include all four scenarios each person encountered, without person fixed effects.

discriminatee race only varies between the experiment's two Stages. In Section 5, we will document and scrutinize these order effects in greater detail, and exploit them to shed light on the tension between two models of fairness (utilitarian versus race-blind rules) among the moderate and liberal respondents to our survey.

3. Some Facts

This Section describes how fairness perceptions in our survey depend on the respondent's personal characteristics; on the experimental treatment the respondent encountered; and on some interactions between these (for example, between the respondent's political orientation and the race of the fictitious discriminatee). As already noted, to avoid any influence of treatment order effects for the *race* treatment, this entire Section uses only responses from Stage 1 of the survey, giving us two responses per subject.[13] To account for within-subject correlation of responses, all standard errors are clustered by subject.

3.1. How Does Perceived Fairness Vary with Respondents' Characteristics?

Appendix 2.3 documents how the mean perceived fairness of discriminatory acts varies with respondents' characteristics. To maximize our sample size for these comparisons, we pool responses across both scenarios each subject encountered in Stage 1, regardless of the treatment that was assigned (*motivation*, *justifiability*, or *race*). In short, Appendix 2.3 shows that our subjects' mean fairness assessments do not vary significantly with their age or race. However, women viewed the discriminatory acts as slightly less fair than men. Somewhat surprisingly (to us), respondents' fairness assessments were *positively* related to their education levels; Appendix 2.4 explores this correlation and argues that higher levels of education mostly reflect a higher 'set point' for all fairness assessments, with more-educated

[13] There is no within-subject variation in the race treatment within Stage 1 (or within Stage 2). All of the comparisons described in this Section continue to apply if we go even further and use only data from the very first of the four scenarios each person encountered, although the standards errors are somewhat higher. See for example Figure A6.1.1, which replicates Figure 1 using first-scenario data only.

individuals rating *all* the scenarios they encounter as more fair than less-educated individuals, regardless of the race of the discriminatee and other respondent characteristics (such as political orientation).[14]

Finally, Appendix 2.3 shows that respondents' political leaning has sizable effects on the perceived fairness of discriminatory acts. Self-described conservative respondents perceive these actions to be fairer than both moderates and liberals (e.g., $p = .000$ for conservatives versus liberals). Mean fairness assessments across U.S. political *party* preferences (e.g., Democrats versus Republicans) exhibit similar patterns, though there is a statistically insignificant non-monotonicity, with Independents being more opposed to discrimination than Democrats. Since the Independent group could include people with both extreme right- and left-wing orientations, we use conservative-liberal leaning rather than party affiliation to categorize respondents' political preferences in the remainder of the paper.[15]

3.2 Treatment Effects on Fairness Assessments

In Figure 1, we compare the fairness assessments of subjects who were exposed to the Tastes versus Statistical treatments, and to the more versus less justifiable forms of each. As in Section 3.1. we pool both of the Stage 1 scenarios encountered by each worker and cluster our standard errors by respondent. To simplify the presentation, we also pool the Black and White treatments.[16] To facilitate interpretation here and throughout the paper, we report all fairness assessments on a scale from -3 ("very unfair") to 3 ("very fair"), where 0 was labeled in the survey as "neither fair nor unfair."[17] The standard deviation of fairness assessments in Stage 1 is 1.657 across respondents, 0.961 within respondents, and 1.915 overall.

According to Figure 1, the average respondent sees no meaningful distinction between the fairness of the statistical versus taste-based scenarios in our survey ($p = .971$). Conditioning on whether

[14] Specifically, Appendix 2.4 shows that educated peoples' higher fairness assessments are not related to differences in political affiliation across education categories: The positive association between education and overall fairness ratings remains very strong within both conservative and liberal survey respondents. Also, more-educated respondents' greater tolerance of the discriminatory acts we depict is not confined to discrimination against a particular race: in fact, the phenomenon applies equally to anti-Black and anti-White discrimination.
[15] All the results by political party are very similar, with occasional non-monotonicities similar to Figure A2.3.1(e), where Independents appear to be to the left of Democrats.
[16] Figure 2 shows that the effects of *justifiability* are virtually identical for White versus Black discriminatees.
[17] As noted, the subjects saw these verbal descriptions, associated with the numerals 1 through 7.

discrimination is taste-based or statistical, however, subjects view the less justifiable form of either taste-based or statistical discrimination as less fair than the more justifiable form (p = .000 in both cases), confirming our expectations. To illustrate the size of these differentials, we first remark that an average respondent did not view the more-justifiable forms of either statistical or taste-based discrimination (high quality information; accommodating the tastes of others) as unfair at all: the mean fairness ratings of these actions were in the "somewhat fair" range with small standard errors.[18] In contrast, the less justifiable forms of taste and statistical discrimination were both viewed much more harshly—specifically 0.925 units (on a scale of -3 to 3), or 0.483 standard deviations less fair.

In Figure 2 we turn our attention to the *race* treatment—i.e. the race of the person who was discriminated against. Motivated by Figure 1 (which shows no difference between the Statistical and Tastes treatments) we now pool these treatments but continue to distinguish between their more- versus less-justifiable forms. In the sample as a whole, Figure 2 shows that respondents view the same discriminatory acts more negatively when they are directed at Black than at White job applicants. These differences are substantial in magnitude, amounting to about 0.5 fairness units or 0.263 standard deviations, and are highly statistically significant (p= .002 and .000 within the *less* versus *more* justifiable forms of discrimination, respectively).

3.3 Discriminatee Race Effects by Respondent Race and Political Orientation

While the effects of the *race* treatment shown in Figure 2 are interesting, these effects may not be the same for all types of respondents. For example, one might expect Black respondents to react more negatively than White respondents to discrimination against Black job applicants. To explore this issue, Figure 3 presents separate estimates of the discriminatee race effect for respondents of different races. Unfortunately, our samples of both Black and Other racial groups are too small to precisely estimate a discriminatee race effect within either group. The point estimates for these groups do however suggest that both groups respond to the race of the discriminatee in much the same way as White respondents

[18] The confidence interval for the fairness of *more*-justifiable Taste-based discrimination includes zero (neither fair nor unfair); for *more*-justifiable Statistical discrimination the confidence interval is bounded above zero.

do.[19] In sum, Figure 3 underscores the fact that the discriminatee race effect in our data – i.e., the tendency to see discrimination against Black people as less acceptable than discrimination against White people—is driven primarily by our White respondents, who comprise about 78 percent of the sample. Thus, while we continue to estimate all our results on the full sample of MTurk respondents in the remainder of the paper, it is important to bear in mind that the stark political differences we will document throughout the paper are driven, to a substantial degree, by differences between White respondents with different political leanings.

Turning to those political differences, Figure 4 presents separate estimates of the discriminatee race effect by the respondent's political leaning. These reveal a clear difference: the discriminatee race effect is stronger among moderate and liberal respondents than in the sample as a whole, but is absent among conservatives. Conservatives view discrimination against (fictitious and identically qualified) Black and White job applicants the same way: as more fair than unfair.[20] A final striking finding from Figure 4 is the strong similarity in both the levels of fairness rankings and in the discriminatee race effects between self-described moderate and liberal respondents. Later in the paper (starting in Section 4.4) we take advantage of this similarity to simplify our analysis by comparing just two political groups—conservatives versus moderates/liberals.

4. Understanding the Facts-- Assessing Four Models of Fairness

This Section describes four simple models of how subjects might evaluate the fairness of discriminatory actions: (simple) utilitarianism, racial in-group bias, race-blind rules (RBR), and belief-based utilitarianism (BBU). For each model, we compare its predictions with the main empirical patterns in our data, and ultimately reject two of the models –racial in-group bias and BBU—as relevant to our context because their predictions are starkly inconsistent with some key facts. We conclude by estimating a regression model that quantifies the relative importance of the two remaining models –utilitarianism and

[19] Interestingly, Figure 3 indicates that the Other group views discrimination relatively harshly. However, there is little indication of a discriminatee race effect for this group of respondents (p = .506) and the point estimates themselves are somewhat imprecise.

[20] Both confidence intervals are bounded above zero ("neither fair nor unfair").

RBRs-- for each of our two political groups. As in Section 3, our analysis only uses data from Stage 1 of the experiment to ensure that *race* treatment order effects cannot affect our conclusions.

4.1 (Simple) Utilitarianism

Utilitarian models of fairness share two main components, the first of which is that fairness depends on outcomes, not on intentions or justifications. In our case, this means that fairness depends on the consequences of the employer's choice – that one applicant got the job offer and the other did not—and not on the reasons why the employer made that choice. Second, utilitarian welfare criteria assign higher levels of fairness to outcomes that redistribute wealth or opportunities from people with higher incomes to those with lower incomes. In the context of our survey, utilitarian respondents should therefore assign lower levels of fairness to acts of discrimination against Black people, whose incomes are on average significantly lower than White peoples'. Utilitarian respondents should not distinguish between Taste-based versus Statistical discrimination, nor among the less or more justifiable forms of each.

We refer to the type of utilitarianism described in this Section as 'simple' because, In addition to reducing income disparities between groups, utilitarians might also wish to equalize other features of the economic environment, including the economic *opportunities* available to racial groups. As we shall document, however, our respondents' perceptions about racial differences in opportunities vary dramatically. (See also Davidai and Walker (2021), Kraus et al. (2017), and Kraus et al (2019) who document substantial misperceptions of racial opportunity gaps.) For this reason, we frame our 'simple' utilitarian model in terms of racial *income* differences, because Black Americans' incomes are indisputably lower. We will address the effects of perceived opportunity gaps on fairness assessments under the heading of *belief-based utilitarianism*. In Section 4.4.

Turning to the evidence on simple utilitarianism, as already shown in Figure 2 respondents in general *do* view discrimination against Black applicants more harshly than discrimination against White applicants. That said, Figure 4 showed that this tendency was confined to moderates and liberals: Conservatives do not consider race when assessing the fairness of discriminatory actions. We conclude that (simple) utilitarian preferences may play a role in moderates' and liberals' fairness assessments, but

are not consistent with conservatives' fairness statements. Utilitarianism also cannot account for the large justifiability effects that are documented in Figures 1 and 2.

4.2 Racial in-group bias

Like utilitarianism, most models of in-group bias are *consequentialist* in nature: they assign fairness to actions based on the action's consequences, not on the actor's intentions.[21] The difference is that –instead of favoring actions that benefit lower-income groups-- persons motivated by racial in-group bias will favor actions that redistribute resources from members of other races to members of their own. Recent evidence of in-group bias includes work by Luttmer (2001), Chen and Li (2009), and Fong and Luttmer (2009, 2011). In our experiment, which studies a survey respondent's assessment of discriminatory acts against a (fictitious) employee, respondents who exhibit racial in-group bias should view those acts as less fair when the fictitious discriminatee shares the respondent's race.[22] As Figure 3 has already shown, we do not have the statistical power to test these predictions for the respondents in our Black or Other racial categories.[23] Our evidence for White respondents, however, is strongly inconsistent with racial in-group bias: As a group, White respondents view discrimination against Black people as substantially *less* fair than discrimination against White people.[24] Overall, we conclude that racial in-group bias model does not provide a useful lens for understanding the main fairness ratings patterns we have documented.

4.3 Race-Blind Rules (RBR)

In contrast to utilitarianism and in-group bias, *rules-based* models of fairness are not consequentialist in nature; instead, they belong to the class of *deontological ethics*, which associate fairness with adherence to a consistent set of rules (Andreoni et al. 2019). Further, in deontological

[21] Chen and Li (2009) review a variety of models that could account for in-group bias in allocation decisions.
[22] Related (and with the same empirical predictions in the case of our experiment) we would also expect respondents to more forgiving of discriminatory acts committed by a member of their own racial group.
[23] In this respect, our MTurk sample is no different from any nationally representative sample of this size. Without quota-sampling minority respondents (which is not possible on MTurk) a much larger sample would be needed to measure the amount of in-group racial bias among other racial groups.
[24] Figure A6.2 replicates Figure 4 (which shows discriminatee race effects by political orientation) for the subset of our respondents who are White. For White conservatives, Figure A6.2 shows weak evidence that is consistent with racial in-group bias: They rate discrimination against Black people as 0.405 units *more* fair than discrimination against Black people, but this difference is not statistically significant at conventional levels ($p = .134$).

ethics, *intentions* can matter and consequences are secondary: for example, ill-intentioned actions that unintentionally produce a good outcome are considered unethical. Intent and motivation play key roles in civil and criminal law, and abundant evidence from behavioral economics shows that people care about intentions when assessing the fairness of many economic actions.[25] Finally, rules-based models of fairness are *race-blind* when the rules that assign fairness to actions do not depend on the races of the people involved.

Applying these ideas to our experiment, an RBR model of fairness would – unlike the previous two models – allow the fairness of a discriminatory action to depend on the intentions behind it: Did the act serve to indulge the employer's personal racial animus, or to protect his business from retaliation by racist customers? Did the employer do his due diligence before relying on statistical information in hiring, or did he a take hearsay-based shortcut? Further, assuming the respondent has an implicit set of rules defining which of the above motivations are fairer than others, she should apply those rules in a race-blind way. A specific type of discriminatory act should be seen as equally fair or unfair, irrespective of the job applicant's race.

The fairness ratings of our respondents are consistent with the use of race-blind rules (RBRs) in at least three ways.[26] First, the effects of our *justifiability* treatments in Figure 1 strongly support the idea that respondents care about the employer's motivation for discriminating against a job applicant. Second and more strikingly, Figure 2 also shows that our respondents penalized the less-justifiable forms of discrimination by the same amount (relative to the more justifiable forms), regardless of the race of the discriminatee (p=.679). Third, a similar test shows that this stability to discriminatee race also applies to the Taste/Statistical fairness differential—it is essentially zero for both Black and White discriminatees.[27]

---

[25] Intentions are relevant to the distinction between first- and second-degree murder, for example. Charness and Levine (2000), Offerman (2002), Abeler et al. (2010) and Breza et al. (2017) document the effects of intentions on peoples' reactions to layoffs, pay reductions, and pay inequality.

[26] In Section 5.2, we will present a third piece of evidence supporting the RBR model that applies only to moderate and liberal respondents. Specifically, we will argue that the order effects for the Black treatment (which are present only for moderate and liberal respondents) suggest that these respondents prefer to maintain a form of consistency across race in their fairness assessments.

[27] Within Black Discriminatees, Tastes-Based scenarios are 0.121 units more fair. Within White Discriminatees, Taste-Based scenarios are 0.138 units less fair. A test for equality of the Tastes vs. Statistical gap between the Black and White treatment yields p = .319.

In all these cases a change in the motivation for an action has the same effect on the action's perceived fairness, regardless of the race of the discriminatee.

A reasonable concern with these striking justifiability results is that they might be driven by the fact that our subjects always encounter both the *less* and *more* justifiable forms, one after the other, within each Stage. We also draw subjects' attention to the differences between the two successive scenarios in the survey instructions. Thus our respondents might have paid close attention to the *relative* fairness ratings they assign to less- versus more-justifiable scenarios. As discussed in Section 6's robustness analysis, we explored this issue by replicating Figures 1 and 2 using only the first scenario each respondent encountered. These estimates use cross-sectional variation only to identify the effects of *justifiability*, and are uncontaminated by any possible desire to be consistent with previous fairness assessments the respondent made. Remarkably, the results are essentially indistinguishable from Table 1. We conclude that subjects' desires to maintain a consistent ranking of these forms of discrimination (perhaps because of experimenter demand effects) are not responsible for these patterns.

Reconciling our justifiability results with a purely consequentialist model of fairness is challenging, in part because the consequences of the employer's discriminatory act – being denied a job offer — are held constant in Figure 1.[28] Still, a consequentialist might argue that the less-justifiable forms of discrimination inflict a greater amount of *psychological* harm on the discriminatee. For example, it might be more painful to be rejected because the boss didn't like your race than because his customers didn't like your race. If so, a utilitarian welfare criterion *could* justify a larger fairness penalty for a poorly justified act of discrimination. That same utilitarian criterion, however, should attach a larger fairness penalty to poorly justified discriminatory acts against Black discriminatees, compared to Black discriminatees (because the acts inflict psychological harm on a lower-income group). As Figure 2 showed, this is not the case: The low- versus high-*justifiability* gaps for White versus Black discriminatees are very similar (-0.953 versus -0.898 fairness points respectively with $p = .679$ for a test

[28] While the material consequences of not being hired could vary with the discriminatee's race (because of differences in outside labor market options), Figure 1 varies only the *reasons* for not being hired: discriminatee race is balanced between the motivation and justifiability treatments due to random assignment. Figure 2 also shows that the justifiability treatments have similar effects even after we condition on discriminatee race.

of equality). Thus, even introducing arbitrary *psychological* consequences that vary with the intent of the act cannot rescue a consequentialist explanation of the *justifiability* treatment effects we observe.

We conclude our discussion of race-blind rules by noting that the preceding evidence in their favor applies on both sides of the U.S. political divide. We show this explicitly in Figure 5, which shows that respondents ranked the relative fairness of *more* versus *less* justifiable forms of discrimination almost identically, irrespective of their political leaning. In sum, there is substantial *prima facie* evidence of deontological ethics based on race-blind rules among our subjects: Subjects care about the reasons why a discriminatory act occurred in a consistent manner (Tastes versus Statistics does not matter; other motivational factors captured by our *justifiability* treatments do matter). Consistent with a widely held desire to adhere to race-blind rules, these motivational factors affect the perceived fairness of a discriminatory action in strikingly similar ways regardless of the race of the discriminatee, and regardless of the political orientation of the respondent.

4.4 Belief-Based Utilitarianism (BBU)

In Section 4.1 we ruled out (simple) utilitarianism among conservative respondents because those respondents did not object more strongly to anti-Black than to anti-White discrimination, despite the fact that Black job applicants, on average, have lower incomes. This fact, however, does not rule out the possibility that conservatives are motivated by a different form of utilitarianism, which we label *belief-based utilitarianism* (BBU).[29] Under BBU, respondents still value redistribution from more- to less-advantaged groups, but they use a different and possibly subjective metric of relative advantage to guide their fairness evaluations.[30] From a modeling perspective, this is an appealing hypothesis because it would allow us to explain a key empirical difference between conservatives and other respondents—conservatives do not exhibit a discriminatee-race effect—in a straightforward way: Both conservatives

[29] BBU is essentially the conceptual framework laid out in Alesina et al. (2020), and underlying the empirical work in Alesina et al. (2021): People have beliefs about the relative incomes and opportunities available to different demographic groups, then use a utilitarian ethic (favoring the lower-opportunity group) to translate these beliefs into support (or non-support) for public policies.

[30] The data in our survey do not allow us to distinguish whether respondents' beliefs about relative opportunities motivate their perceptions of the fairness of discriminatory acts, or whether these beliefs are *motivated* by a desire to evaluate discriminatory actions in a certain way. Oprea and Yuksel (2021) use a cleverly designed experiment to detect motivated beliefs in a different context from ours.

and other respondents are in fact utilitarians (i.e. they prefer to favor a disadvantaged group) but they simply have different beliefs about who is disadvantaged.

Evidence that appears to support BBU is presented in Figure 6, which draws on our survey's measure of Black people's relative economic opportunities (BRO-- from the common questions at the end of the survey). This question asked the respondents to rate Black people's relative economic opportunities in the United States on a seven-point scale, running from "much less opportunity" to "much more opportunity". Figure 6 shows that the respondents' BRO ratings differ dramatically by their political orientation: While liberals have a mean BRO of -1.374 (p = .000), conservatives' mean of -0.206 is insignificantly different from zero (p = .089) with moderates in between. This suggests that conservatives' belief that Black and White people have roughly equal opportunities has the potential to account for their observed fairness ratings, which –like their BRO ratings—are statistically the same for discrimination against Black versus White job applicants.[31]

To assess whether BRO differences can actually account for the partisan gap in fairness assessments, panel (a) of Figure 7 shows respondents' fairness ratings for discrimination against Black applicants by BRO categories, separately for conservatives and moderates/liberals. If BRO accounts for the large partisan gap, we should see little or no partisan gap *within* the BRO categories; the partisan gap should be explained, instead, by the higher mean level of BRO among conservatives. The evidence, however, paints a very different picture in two key respects. First, while BRO is very predictive of the perceived fairness of anti-Black discrimination among *moderates and liberals*, it is not predictive of conservatives' fairness ratings. In other words, we see no effect of BRO on perceived fairness of anti-Black discrimination among conservatives, even though their beliefs about relative racial opportunities vary widely. Second, Figure 7 shows that there are large political gaps in the perceived fairness of discriminating against Black people, even when we condition on BRO. These political gaps are

[31] Our findings about the partisan gap in perceived relative opportunities (BRO) mirror the partisan differences in perceptions about inequality and mobility documented by Alesina et al. (2020), and the stark partisan differences in beliefs about the causes of racial inequality documented by Alesina et al. (2021). They also mirror Alesina et al.'s (2021) and Haaland and Roth's (2021) findings that Democrats perceive that there is much more anti-Black discrimination than Republicans do. As noted, our contribution relative to these papers is that we study the fairness of individual (discriminatory) actions (not public policies) and we demonstrate the key role of the intentions behind an action in determining its perceived fairness.

particularly stark at the bottom of the BRO distribution: While moderates and liberals who think that "Black people have much less opportunity than White people" (BRO=-3) are strongly opposed to anti-Black discrimination, conservatives with the same belief are, on average, *accepting* of anti-Black discrimination (with a mean fairness rating of about +0.5). [32]

A third and even more surprising piece of evidence against the "BRO hypothesis" emerges from panel (b) of Figure 7, which replicates panel (a) for discrimination against White job applicants. Consistent with a large explanatory role for BRO in peoples' fairness assessments, this Figure shows an effect of BRO on the perceived fairness of discrimination that is essentially invariant to political orientation: the coefficients are .257 (p = .094) and .265 (p=.000) for conservatives and moderates/liberals respectively. However, for both political groups the direction of this effect is the opposite of what the BRO hypothesis would predict: According to the BRO hypothesis, higher levels of Black people's perceived relative opportunities should make discrimination against White people less acceptable. Instead, the perception that Black people have equal or more economic opportunities than White people – which is held by 36.9 percent of our subjects—is associated with a *greater* tolerance of (hypothetical) acts of anti-White discrimination.

Summing up, while respondents' stated beliefs about Black peoples' relative opportunities (BRO) are are (a) correlated with their political affiliations and (b) quite predictive of respondents' assessments of the fairness of discriminatory acts, the signs and patterns of these associations are decidedly inconsistent with the 'BRO hypothesis': the idea that conservatives' beliefs about Black relative opportunities explain their tolerance of anti-Black discrimination. Instead, it appears that high levels of BRO may be better understood as a marker of a political attitude that is accepting of discriminatory acts, *regardless of the recipient of the act*. While this attitude is more common among conservatives than other political groups, Figure 7 shows that it is held by some members of all political groups.[33]

---

[32] This partisan gap at the bottom of the BRO distribution is highly statistically significant. Both within subjects who have BRO levels of -3, and within subjects who have BRO levels of -2, the partisan gap is significant at p=.000.

[33] This result complements Haaland and Roth (2021), who randomly administer an information treatment that successfully reduces the partisan disparity in beliefs about Black people's relative economic opportunities. This induced change in beliefs, however, does not reduce the partisan gap in their subjects' support for pro-Black economic policies.

4.5 Quantifying the Relative Roles of Utilitarianism and RBR preferences

Motivated by the lack of support for racial in-group bias and BBU in our data, in this Section we estimate a simple regression model of respondents' fairness assessments that allows respondents to care about just two forms of fairness: utilitarianism and race-blind rules (RBR). In this model, respondents' fairness assessments, $Fair_{it}$, can be written as:

$$Fair_{it} = \alpha + \beta_1 Low_{it} + \beta_2 Black_{it} + \beta_3 (Low_{it} * Black_{it}) + \epsilon_{it} \quad (1)$$

where Low_{it} indicates exposure to a low-*justifiability* scenario and $Black_{it}$ indicates that respondent i was exposed to a Black discriminatee in the t^{th} scenario i encountered. In equation (1), β_1 (the additional fairness associated with low-*justifiability* acts of discrimination) quantifies respondents' concerns with the discriminator's *intentions* (when the discriminatee is White); β_2 (the additional fairness associated with a Black discriminatee) quantifies the strength of utilitarian preferences (in high-justifiability scenarios). Finally, β_3 tests the extent to which subjects' responses to the discriminator's intentions are race neutral: If $\beta_3 = 0$, respondents attach the same fairness penalty to a low-*justifiability* action, regardless of the racial identities of the people involved. Notice that equation (1) allows individual respondents to care about both race-blind rules and utilitarian concerns, and allows us to measure the relative strength of those concerns. Motivated by the robust lack of fairness differentials between Taste-based and Statistical discrimination, equation 1 does not distinguish between these forms of discrimination to simplify the presentation.

Estimates of equation (1) are provided in Table 1, separately for the sample as a whole, for conservative respondents, and for moderate plus liberal respondents. Like all the analyses in Sections 3 and 4, Table 1 uses data from Stage 1 of the survey only; thus, we have two observations per respondent ($t \in (1,2)$). Focusing first on conservatives, column 3 of Table 1 –which omits the interaction term to give us summary measures of overall effects-- shows that these respondents use some rules-based ethics: compared to more-justifiable discriminatory acts, less-justifiable acts are 0.865 units (or 0.452 standard deviations) less fair. In contrast, there is no evidence for any utilitarian concerns among conservatives: The race of the discriminatee has no detectable effect on conservatives' fairness assessments. Finally,

although the standard error is high, the small point estimates of β_3 (the interaction term) in column 4 indicate that conservatives implement their personal fairness rules in a race blind way.

Turning to moderates and liberals, column 6 of Table 1 shows that they also apply rules-based ethics: compared to more justifiable discriminatory acts, less-justifiable acts are 0.947 units (or 0.495 standard deviations) less fair. This penalty is very similar to conservatives', and the two estimates are statistically indistinguishable (p = .600). In stark contrast to conservatives, however, liberals exhibit strong utilitarian preferences: The same act of discrimination is 0.790 units (or 0.413 standard deviations) less fair when it is directed at a Black than a White job applicant. Notably, as defined here, these measures of the strength of moderates/liberals' utilitarian versus RBR preferences are roughly equal in magnitude.[34] Finally, the point estimate of β_3 in column 6 indicates that –like conservatives-- liberals and moderates implement their fairness rules (with respect to the nature of the discriminatory act) in a race blind way.[35] As noted earlier, the similar way in which the two political groups react to the specific circumstances of a discriminatory act (i.e. $\beta_1$and β_3 in equation 1), appears to describe some important common ground in these groups' beliefs about what is fair in the labor market.

5. Learning from Order Effects: Experimenter Demand Effects and the Utilitarianism-RBR Tradeoff

In Section 4.5 we argued that –while conservatives' fairness assessments appear to be influenced only by race-blind rules (RBRs)—moderates and liberals seem to care about both RBRs and utilitarian criteria. In Section 2.4 we showed that subjects' fairness assessments exhibit order effects: their Stage 2 ratings depend on race of the discriminatee they encountered in Stage 1. In this Section, we exploit these order effects to study how liberals and moderates reconcile the two fairness criteria they care about – utilitarianism and RBRs—when those criteria conflict.

[34] We recognize that comparing discriminatory acts of different justifiability versus acts against different races is not a natural metric; the main goal of the Table 1 regressions is to quantify respondents' reactions to two different fairness determinants in a multiple regression context that allows both to operate simultaneously.
[35] A test of equality for the estimates of β_3 between both groups is also statistically indistinguishable (p = .447).

To accomplish this goal, we proceed in three steps. First, we exploit data from Stage 3 of the experiment to rule out a plausible form of experimenter demand effects as an explanation for the order effects we observe. Second, we document that these order effects are only present among moderate and liberal respondents. Finally, we use a simple model of reporting behavior, combined with random assignment of the *race* treatment to interpret moderates and liberals' order effects as a compromise between utilitarianism versus RBRs, and to estimate the relative weight moderates and liberals place on those two criteria when they conflict. The key assumption underlying our approach is that respondents only become conscious of their desire to be race-blind when they encounter discriminatees from a second racial group.

5.1 Experimenter Demand Effects

Intuitively, the order effects we observe are that subjects have a less negative view of anti-Black discrimination in Stage 2 if they encountered a White discriminatee instead of a Black discriminatee in Stage 1.[36] One potential explanation of such a response is an experimenter demand effect of the following form: If respondents encounter the Black treatment in Stage 1, they assume that we (the experimenters) are either moderate or liberal. Then --to please us-- the respondents provide Stage 1 fairness assessments that are typical for moderates and liberals (i.e. discrimination against Black applicants is unfair, and more unfair than discrimination against White applicants). On the other hand, if respondents encounter the White treatment in Stage 1, they assume the experimenters are conservative and provide Stage 1 answers that are typical for conservatives (i.e., discrimination against both Black and White applicants is neutral or fair). Finally, respondents who encounter a change in the *race* treatment between Stages 1 and 2 update their priors to become uncertain about the experimenters' politics and moderate their fairness assessments accordingly.

To probe the plausibility of this demand-effects model, Appendix 4 argues that subjects who want to please the experimenters should tailor not just their fairness assessments but also their answers to other survey questions to achieve the same end. Of particular interest in this regard are the subjects'

[36] We also find that subjects' Stage 2 assessments of anti-White discrimination are more negative if they encountered a Black discriminatee in Stage 1, although this difference is not statistically significant.

assessments of Black peoples' relative economic opportunities (BRO), and potentially even subjects' reported political orientations (all elicited in Stage 3 of the survey). For example, suppose a subject encountered the White treatment in both Stage 1 and 2 of the survey. Under our assumptions about experimenter demand effects, this should send a strong signal that the experimenters are conservatives. To please us, we would then expect the subjects to report that Black people have a higher level of relative economic opportunity and perhaps even to report their own political leanings as more conservative.

In Appendix 4, we examine whether subjects' responses to these Stage 3 questions depend on the *race* treatments they received in Stages 1 and 2, and find no such effects: Specifically, subjects' BRO assessments, stated political party preferences, and reported left-right leaning are highly stable with respect to the *race* treatments they encountered earlier in the experiment. We conclude that experimenter demand effects of this type are probably not responsible for the order effects we observe.

5.2 *Race* Treatment Order Effects are Absent among Conservatives

Taking it as given that our respondents' political learnings are honestly reported, we now establish one additional fact concerning the *race* treatment order effects in our data: these order effects are absent among conservatives. Specifically, Appendix 5 replicates Figure A3.4.1 –which showed that respondents' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1— separately for conservative respondents versus moderate / liberal respondents. It shows that there are no such order effects for conservatives: Regardless of discriminatee race they encountered in Stage 1, conservatives view discrimination as a little more fair than unfair (about +0.5 on a scale from -3 to +3) in Stage 2. Moderate and liberal respondents, on the other hand, exhibit a more pronounced version of the aggregate order effects documented in Figure A3.4.1: Moderates' and liberals' Stage 2 fairness assessment of anti-Black discrimination are much milder if they encountered a White discriminatee (as compared to a Black discriminatee) in Stage 1 $p = .009$).[37] Motivated by this fact, we restrict our attention to moderate and liberal respondents in Section 5.3, where we propose an interpretation of the *race* treatment order effects in our data.

[37] The race treatment a moderate/liberal subject received in Stage 1 does not have a statistically significant effect on the subject's ratings of discrimination against White people in Stage 2.

We conclude by noting an additional implication of the fact that order effects are absent among conservatives. Specifically, this fact makes the "experimenter demand effects" hypothesis examined in Section 5.1 even less plausible. For experimenter demand effects to explain our results, these demand effects must *only* be present among moderates and liberals. In other words, moderates and liberals should want to please an experimenter they perceive as moderate or liberal, but conservatives must have no such desire to please a conservative experimenter. In sum, experimenter demand effects would need to take a very special form --only affecting one of our Section 3 questions about beliefs and political attitudes, and only affecting one political group-- to account for the order effects in our data.[38]

5.3 A Trade-off between Utilitarianism and Race-Blind Rules?

In Section 4.5, we used Stage 1 data to argue that moderate and liberal subjects may have utilitarian motives, but also appear to derive utility from race-blindness when making fairness assessments. If that is the case, subjects who are exposed to both Black and White treatments (i.e. treatment switchers) face a conflict between utilitarianism and race-blindness. For example, in Stage 2, a White-to-Black treatment switcher needs to choose between assigning the same fairness rating they assigned to a White discriminatee in Stage 1 (race blindness), and respecting their utilitarian desire to object more strenuously to anti-Black than anti-White discrimination. Subjects who do not experience *race* treatment changes do not face this conflict.

To model this idea, we make the following assumptions:

Assumption 1:

Subjects' Stage 1 assessments, B_i^1 and W_i^1represent each respondent *i*'s "true" ratings in a setting where they don't need to consider race-blindness (B_i^* and W_i^*).

[38] In contrast, note that the 'tradeoff' model in Section 5.3 has a 'built-in' explanation for the absence of order effects among conservatives: Because conservatives exhibit no utilitarian concerns (as is demonstrated by their Stage 1 reports) conservatives *never face a conflict* between utilitarian preferences and a desire to be race neutral. Thus, the model itself predicts that order effects should only be present among groups (like moderates and liberals) whose Stage 1 responses indicate that they care about fairness criteria *other* than RBRs.

Assumption 1 seems reasonable because in Stage 1, respondents have not been asked to make any previous fairness assessments with which they might want to be consistent.

Assumption 2:

In Stage 2, *race* treatment switchers care about two potentially conflicting things: reporting their true rating and making the same report as in Stage 1 (being race-blind).[39]

Using this notation, White-to-Black treatment switchers have the option of reporting their true rating of discrimination against the second racial group they encounter ($B_i^2 = B_i^*$), assigning the same rating they assigned (to the other race) in Stage 1 ($B_i^2 = W_i^1$), or reporting a weighted average of these two choices:

$$B_i^2 = \alpha B_i^* + (1-\alpha) W_i^1 \quad (2)$$

where α is the weight placed on their 'true' utilitarian preference and (1- α) is the weight on their desire to make race-blind assessments. Our goal is to estimate α, but this is complicated by the fact that (unlike W_i^1 and B_i^2), B_i^* is not observed for White-to-Black treatment switchers.

To address this unobservability problem we take advantage of the fact our *race* treatments are randomly assigned. Thus, while B_i^* is not observed for W-to-B switchers (and W_i^* is not observed for B-to-W switchers), their sample means $\bar{B}^*$ and $\bar{W}^*$ in any fixed population (such as moderates and liberals) *are* observed for both groups of switchers from the mean Stage 1 responses of the subjects in their population who were randomly assigned to the other *race* treatment. We can therefore write:

$$\bar{B}^2 = \alpha \bar{B}^* + (1-\alpha) \bar{W}^* \quad (3)$$

where $\bar{B}^*$ and $\bar{W}^*$ are sample means calculated from Stage 1 responses.

Similarly, for B-to-W switchers,

[39] Subjects' exposure to the Taste and Statistical treatments can change between Stages 1 and 2, but we abstract from that here since those treatments are randomly assigned and never appear to affect fairness assessments.

$$\bar{W}^2 = \alpha\bar{W}^* + (1-\alpha)\bar{B}^* \quad (4)$$

After restricting the sample to moderate and liberal respondents, Equations (3) and (4) can then be (separately) solved for α, yielding $\alpha = 0.44$ for the White-to-Black switchers and $\alpha = 0.62$ for the Black-to-White switchers.[40] Thus, W-to-B switchers' Stage 2 ratings of anti-Black discrimination are (slightly) closer to strict race-blindness ($\alpha = 0$) than to a pure utilitarian assessment ($\alpha = 1$). B-to-W switchers, on the other hand, act as if they place slightly more weight on their utilitarian 'truth' than on race-blindness. That said, we cannot reject equal weight on both objectives ($\alpha = 0.5$) for either type of switcher (p = .678 and p = .423 for W-to-B and B-to-W switchers, respectively).[41]

Additional hypothesis tests allow us to reject other meaningful values of α, however. Specifically, for W-to-B switchers we can reject both $\alpha = 0$ (100% weight on race-blindness) and $\alpha = 1$ (100% weight on the utilitarian 'truth') [p=.003 and p=.007 respectively]. In other words, when moderates and liberals encounter White and Black discriminatees in that order, their second fairness assessment places strictly positive weight on *both* utilitarian preferences and race-blind rules. For B-to-W switchers we can reject pure race-blindness (p=.000), but we cannot quite reject 100% weight on utilitarian preferences (p=.067). In sum, the *race* treatment order effects we observe among our moderate and liberal respondents can be explained by a simple model that assumes these respondents value both the race-blind application of rules (RBRs) and utilitarian objectives. When these criteria conflict, i.e. when the respondent experiences a switch in the *race* treatment, respondents roughly 'split the difference' between these two objectives when making their fairness assessments.

6. Robustness

One striking result of our analysis is the large magnitude, statistical significance, and stability of the *justifiability* treatments: Respondents of all political orientations penalized the less- justifiable forms

[40] Our lower estimate of α (the weight placed on the utilitarian 'truth') for W-to-B switchers than B-to-W switchers is consistent with the fact that *race* treatment order effects were only statistically significant for the former group.
[41] The 95% percent confidence intervals for α are $[0.217,0.706]$ and $[0.171,0.644]$ for W-to-B and B-to-W, respectively.

of discrimination (relative to more-justifiable forms) by the same amount, irrespective of the discriminatee's gender. A possible concern with this result is the fact that the subjects always encounter both the *less* and *more* justifiable forms within each Stage (one right after the other), and that we draw subjects' attention to the sentences in the two scenarios that differ from each other. Thus, subjects may have taken special care to ensure they assign lower fairness rankings to the scenario that feels less justifiable. As noted, Appendix A6.1 addresses this issue by replicating Tables 1 and 2 using only the first scenario each respondent encountered. The results are almost identical to our main estimates, suggesting that subjects' desires to maintain a consistent ranking of the two types of scenarios are not responsible for this finding.

Figure 3 of the paper showed that on average, White respondents object more to acts of anti-Black than anti-White discrimination, and argued on that basis that racial in-group bias does not play a major role in explaining our subjects' fairness assessments. Given the large partisan divide in our data, however, this raises the question of whether racial in-group bias may still be present among a subset of White respondents. To address this issue, Appendix 6.2 replicates Figure 3 for conservative respondents only. Interestingly, the discriminatee race effect *does* switch sign in this group, relative to the full sample in Figure 3: conservative White respondents rate discrimination against Black people as *more* fair than discrimination against White people. This discriminatee race effect is not significantly different from zero at conventional levels, however (p=0.134).

For much of the analysis in the paper we combined moderates and liberals into a single group given their similar levels and patterns of fairness assessments. This includes Figure 7, which illustrated some strong and unexpected relationships between perceptions of Black peoples' relative economic opportunities (BRO), political orientation, and fairness assessments: For example, the acceptability of anti-White discrimination *increases* with Black relative opportunities. To see if the similarity between moderates and liberals extends to these unexpected findings, Figure A6.3 replicates Figure 7, showing separate results for moderates versus liberals. Consistent with earlier results, these two groups exhibit very similar response patterns, both of them differing substantially from conservatives' patterns.

In Section 2.3 we showed that our sample of MTurk respondents was not representative of adult Americans on a number of key dimensions. While our small sample size limits what we can do to address this issue, Appendices 7 and 8 replicate all our main analyses (Figures 1-7 and Table 1) two different ways. First, Appendix 7 uses the 2019 American Community Survey to re-weight our MTurk responses by the relative prevalence of our respondents in 24 cells, defined by gender, race , education, and age. All the main patterns discussed in the paper are replicated, with one small exception: the weak positive association between BRO and the fairness of anti-Black discrimination among conservative respondents in Figure 7(a) becomes somewhat stronger and statistically significant. Similar to Figure 7, however, the slope for conservatives remains much lower than the slope for moderates / liberals.

Second, Appendix 8 replicates Figures 1-7 and Table 1 using weights derived from the 2020 General Social Survey (GSS) which are based only on a 7-point political leaning scale (i.e., extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, and extremely liberal) that is asked in a very similar way to our survey.[42] Despite significant differences in the political mix of the two surveys, all the main results are replicated.[43]

An important concern in virtually all tests of statistical hypotheses is the extent to which the hypotheses were selected after a preliminary analysis of the data. To address this issue, we posted a registered pre-analysis plan (PAP) before launching our survey. The relationships between the analyses proposed in the PAP and the hypotheses tested in our survey are described in detail in Appendix P of the paper. Briefly, Appendices P1-P3 together comprise a "populated PAP" which reports the results of the exact tests specified in the PAP. Appendix P4 then summarizes the relationship between the PAP and the paper.

Specifically, Appendix P4 begins by documenting that the following key analyses in the paper were declared in advance: all the descriptive "facts" presented in Section 3; all four theoretical models of discrimination described in Section 4 and the main tests thereof (the models' names have changed

[42] Because of the small size of the MTurk and GSS samples, we did not re-weight our MTurk sample to mimic GSS demographic characteristics; attempts to do this yielded highly extreme and imprecise weights. The ACS does not ask questions about political orientation or party preference.
[43] The one exception noted with the ACS weights in Appendix 7 does not occur here.

slightly); the possibility of question order effects (especially for the *race* treatment); *and* the idea of using question order effects to learn about respondents' preferences for race-blindness (see Appendix P2.5). Appendix P4 also describes the five most important ways in which our main analyses in the paper differ from the PAP: First to simplify the presentation, the paper mostly reports simple *t*-tests of differences in means rather than regression results. Second, we decided not to standardize our main outcome measure (fairness) because we realized that this would obscure important cardinal information. Third, motivated by the *race* treatment order effects (which we anticipated *might* be present), we restricted the sample in Sections 3 and 4 to Stage 1 responses only. Fourth, we confined Section 5.3's "learning from order effects" analysis to moderates and liberals, because these effects were strikingly absent among conservatives. Fifth, in Figure 7's exploration of the "BRO hypothesis" we decided to use a continuous version of BRO (all seven values) rather than a dichotomized version, to show additional detail.

Finally, Appendix P4 lists the following PAP hypothesis tests that we decided *not* to include in the main paper.[44] We did not include an "actions versus identity" decomposition of fairness determinants because it seemed of limited interest; due to a lack of statistical power we did not pursue the idea of classifying subjects into types based on their responses to within-subject treatment variation; and we did not extend Appendix P3.2's replication of populated PAP results for a subset of 'thoughtful' respondents (who took more than the median time to complete the survey) to the results in the paper. This restriction had very little effect in the populated PAP so we decided the extension would add little to the paper.

7. Discussion

Inspired by a rapidly growing literature on the perceived fairness of pay and income inequality, and by a large literature on discrimination, we have used an MTurk survey to elicit Americans' assessments of the fairness of canonical examples of statistical and tasted-based racial discrimination. We have found, first of all, that respondents of all political leanings are indifferent to the distinction between statistical and taste-based discrimination. This may be surprising given many economists' apparent interests in

[44] The results of these tests are available in Appendices P1-P3.

distinguishing between these types of discrimination.[45] Second, we find that respondents of all political leanings *do* care about other aspects of the motivation behind a discriminatory act—specifically whether statistical discrimination is based on precise information versus hearsay, and whether the discriminator is accommodating his own tastes or those of others. Interestingly, respondents of all political leanings care about these motivational differences in very similar ways, and do so in a race-blind manner, revealing a significant domain of shared values across the political spectrum. Third, there is however a large political difference in both (a) how much respondents object to discriminatory acts in general, and (b) the extent to which their fairness ratings depend on the discriminatee's race. Conservatives do not differentiate between anti-Black and anti-White discriminatory acts, and rate all types of discriminatory acts we depict as either neutral or slightly *more* than fair. Moderates and liberals, on the other hand, view discrimination as unfair in most cases, and object more strongly to anti-Black than anti-White discrimination.

To attempt to understand these facts, we evaluated four pre-specified models of fairness, and showed that two of them –racial in-group bias and beliefs-based utilitarianism—are inconsistent with major features of our data. We conclude, instead, that a model with two political groups and two models of fairness --utilitarianism and race-blind rules-- accounts for most of the response patterns we have identified. In this model, conservatives care *only* about race-blind rules, while moderates and liberals care about both RBRs and utilitarian ethics. When these ethics conflict (e.g. when a subject encounters a switch in the *race* treatment) moderates and liberals assign fairness ratings that place roughly equal weight on each one.

While our main objective in this paper has been to understand when and why people view discriminatory actions as fair or unfair, our findings may also have some implications for both managerial and public policy. In a management or human resources context, our findings suggest that workers' perceptions of the fairness of policies or actions with disparate impacts on racial groups are likely to depend on the precise motivations or circumstances surrounding those policies or actions. In this sense

[45] One motivation for economists' interest in this question may be a perception that statistical discrimination is, indeed, less objectionable than taste-based discrimination. According to this point of view, it seems unfair to blame animus-free employers from simply using available information to hire the best worker. (See Tilcsik 2021 for quotes to this effect from well-known economists). Our results in this paper suggest that the U.S. general population may not share economists' perception about 'tastes' versus 'statistics'.

our findings complement existing evidence that the motivations behind underlying pay differentials (Frank, 1984; Charness and Kuhn, 2007; Gartenberg and Wulf, 2017; Mas, 2017; Breza et al. 2017) and layoffs (Charness and Levine, 2000) have a large effect on their acceptability to workers. Interestingly, since our data show that 'reasons matter' to members of all political groups, our evidence suggests that employers may reap wide benefits from transparent, rules-based recruitment and pay policies that provide clear justifications for any decisions that have disparate racial impacts.

In terms of public policy, our study suggests that motivations could affect the extent to which *public policies* are seen as fair also. If this is the case, winning support for a policy involves not only crafting the details of the policy itself (who wins, who loses) but also crafting a *motivation* for the policy change that is widely perceived as fair. In addition, our results –like Alesina et al.'s (2021) and Haaland and Roth's (2021)-- suggest the potential for substantial political headwinds for certain anti-discrimination policies. While acts of anti-Black discrimination are viewed as unfair by a majority (63.1%) of our sample, the rest of our respondents view the discriminatory actions depicted in our scenarios as either neutral or fair, regardless of the race of the discriminatee. This group of respondents is likely to resist policies that interfere with employers' decisions to hire and fire at will, even when those hiring decisions represent canonical examples of taste-based and statistical discrimination on the basis of race. That said, our analysis also suggests two types of situations in which conservatives might be more receptive to policies that equalize racial opportunities. One of these are cases in which a clear rule has *not* been applied in a race-blind way; in these cases, *restoring* race-blindness should appeal to people with an RBR ethic. Second, our results suggest that conservatives, like moderates and liberals, react more negatively to race-based actions that were taken for less-justifiable reasons, like personal animus and hearsay-based evidence. Antidiscrimination policies that target these types of behaviors may thus be better received by conservatives.

Our analysis is subject to some important *caveats* and limitations. For example, all our results – including our finding that conservatives do not, on average, object to discriminatory actions-- are limited to the range of actions our scenarios depict. It seems likely that more *consequential* acts (e.g. being fired rather than not being hired, or receiving a long prison sentence) or less *justifiable* acts (e.g. ones inspired by racial hatred) would elicit stronger negative responses among all groups. Second, while we have more

than enough statistical power to test our pre-registered hypotheses, we lack the statistical power to resolve some important questions, such as whether self-identified conservatives are more accepting of anti-Black than anti-White discrimination. Constraints on sample size have also prevented us from painting a very detailed picture of the fairness assessments made by non-White Americans.

We also hasten to point out that our analysis has identified some interesting puzzles that are not resolved here. For example, while we demonstrate that race-blind rules can account for several key features of conservatives' fairness ratings, conservatives' reliance on RBRs does not explain why, on average, conservatives do not object to any of the discriminatory acts we depict. For example, instead of *never* penalizing discriminatory acts, an alternative set of race-blind rules would penalize both anti-White and anti-Black discrimination, strongly and equally. We also do not have a good understanding of why subjects of all political leanings become more accepting of (hypothetical acts of) anti-White discrimination as their perception of White peoples' relative economic opportunities becomes more pessimistic. One hypothesis in this regard is that high levels of BRO (Black people's perceived relative opportunities) are a marker for a political attitude that, for example, "too much attention is given to discrimination", and that private discriminatory acts –at least within the range depicted by our scenarios—should simply be tolerated.

Taken together, the preceding *caveats* and puzzles emphasize the fact that our analysis has only scratched the surface of the question that motivates our paper, "When is Discrimination Unfair?". Indeed, we believe that the methods in our paper could easily, and fruitfully, be applied to a wide variety of related questions. For example, discriminatory scenarios in future papers could (a) depict more *consequential* acts (such as getting a long prison sentence); (b) depict less *justifiable* motivations (such as hatred versus mild distaste); (c) vary the *context* in which the act occurs (housing markets, credit markets, judicial decisions); (d) vary the *groups* involved (such as gender, age, sexual orientation, political orientation, age, criminal and credit history); (e) change the *social environment* depicted in the scenario (for example, is the act seen by others?); (f) manipulate whether the fictitious act is *conscious* or unintended; or (g) change the respondent's *decision environment* (priming, mental bandwidth available, and whether the respondent's ratings are visible to others). Our results also suggest that economists' frequent focus on distinguishing taste-based from statistical explanations may misplaced. Instead, it may

be more useful to focus future research more on the aspects of discriminatory acts which our results suggest are of greater concern to most people, such as the detailed motivations behind the acts and the identities of the discriminator and discriminatee.

References

Abeler, Johannes, Sebastian Kube, Steffen Altmann, and Matthias Wibral. 2010. "Gift Exchange and Workers' Fairness Concerns: When Equality is Unfair." *Journal of the European Economic Association* 8 (6): 1299-1324.

Alesina, Alberto, and Eliana La Ferrara. 2005. "Preferences for Redistribution in the Land of Opportunities." *Journal of Public Economics* 89: 897-931.

Alesina, Alberto, Armando Miano and Stefanie Stantcheva. 2020 "The Polarization of Reality" *American Economic Review Papers and Proceedings* 110: 324-328

Alesina, Alberto, Matteo F. Ferroni, and Stefanie Stantcheva. 2021. "Perceptions Of Racial Gaps, Their Causes, And Ways To Reduce Them" NBER working paper no. 29245

Almås, Ingvild & Cappelen, Alexander & Tungodden, Bertil. (2020). "Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?" *Journal of Political Economy*, Volume 128, Number 5.

Andreoni, James, Deniz Aydin, Blake Barton, B. Douglas Bernheim and Jeffrey Naecker. 2020. "When Fair Isn't Fair: Understanding Choice Reversals Involving Social Preferences." *Journal of Political Economy* 128(5): 1673-1711.

Arechar, Antonio A. Simon Gachter, and Lucas Molleman. 2018. "Conducting Interactive Experiments Online." *Experimental Economics* 21: 99-131.

Auspurg, Katrin, Thomas Hinz, and Karsten Sauer. 2017 "Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments" *American Sociological Review* Vol 82, Issue 1.

Barr, Abigail, Tom Lane and Daniele Nosenzo. 2018. "On the Social Inappropriateness of Discrimination." *Journal of Public Economics* 164: 153–164.

Bohren, J. Aislin, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." NBER Working Paper No. 25935.

Bracha, Anat, Uri Gneezy, and George Loewenstein. 2015. **"**Relative Pay and Labor Supply." *Journal of Labor Economics* 33 (2): 297-315.

Breza, Emily, Supreet Kaur, and Yogita Shamdasani. 2017. "The morale effects of pay inequality." *The Quarterly Journal of Economics 133(2)*: 611-663.

Bruhin, Adrian, Ernst Fehr, and Daniel Schunk. 2019. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association* 17(4): 1025–1069.

Cain, Glen G. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey." In *Handbook of Labor Economics*, Vol. 1, edited by O. Ashenfelter & R. Layard, 693-781. Elsevier.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. "Inequality at work: The effect of peer salaries on job satisfaction." *American Economic Review* 102(6): 2981-3003.

Charness, Gary., Till Gross, and Christopher Guo. 2015. "Merit Pay and Wage Compression with Productivity Differences and Uncertainty." *Journal of Economic Behavior & Organization* 117: 233-247.

Charness, Gary and David I. Levine. 2000. "When Are Layoffs Acceptable? Evidence from a Quasi-Experiment." *Industrial and Labor Relations Review* 53(3): 381-400.

Charness Gary and Peter Kuhn. 2007. "Does Pay Inequality Affect Worker Effort? Experimental Evidence". *Journal of Labor Economics* 25(4): 693-724.

Chen, Y. and Li, S. X. 2009. "Group identity and social preferences." *American Economic Review* 99(1): 431–457.

Cohn, Alain, Ernst Fehr and Lorenze Götte. 2014. "Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment." *Management Science*, 61(8): 1777-1794.

Cullen, Zoe B. and Bobak Pakzad-Hurson. 2017. "Equilibrium Effects of Pay Transparency." unpublished paper, Harvard Business School.

Davidai, S. and J. Walker (2021). Americans Misperceive Racial Disparities in Economic Mobility. *Personality and Social Psychology Bulletin*, 01461672211024115.

Feess, E., Feld, J., and Noy, S. (2021). "People Judge Discrimination Against Women More Harshly Than Discrimination Against Men - Does Statistical Fairness Discrimination Explain Why?" *Frontiers in psychology*, *12*, 675776.

Fehr, Dietman, Hannes Rau, Stefan T. Trautmann and Yilong Xu. 2021. “Fairness Properties of Compensation Schemes”. Unpublished paper, University of Heidelberg.

Fong, C. M. and E. F. Luttmer (2009). What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty. *American Economic Journal: Applied Economics* 1 (2), 64-87

Fong, C. M. and E. F. Luttmer (2011). “Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment” *Journal of Public Economics* 95 (5), 372-394.

Frank, Robert H. 1984. “Are Workers Paid Their Marginal Products?” *American Economic Review* 74(4): 549–571.

Gartenberg, Claudine, and Julie Wulf. 2017.“Pay Harmony? Social Comparison and Performance Compensation in Multibusiness Firms.” *Organization Science* Vol. 28(1): 39-55.

Griggs v. Duke Power Co. 1971. 401 U.S. 424.

Haaland, I. and C. Roth (2021). Beliefs about racial discrimination and support for pro-black policies. *Review of Economics and Statistics*, forthcoming.

Jasso, Guillermina and Peter H. Rossi (1977) Distributive Justice and Earned Income *American Sociological Review* Vol. 42, No. 4 (Aug. 1977), pp. 639-65.

Jasso, Guillermina, Robert Shelly and Murry Webster 2019 “How impartial are the observers of justice theory?” *Social Science Research* 79: 226-246.

Kraus, M. W., I. N. Onyeador, N. M. Daumeyer, J. M. Rucker, and J. A. Richeson (2019). The Misperception of Racial Economic Inequality. *Perspectives on Psychological Science* 14 (6), 899–921.

Kraus, M. W., J. M. Rucker, and J. A. Richeson (2017). Americans misperceive racial economic equality. *Proceedings of the National Academy of Sciences* 114 (39), 10324–10331.

Kuhn, Peter and Trevor Osaki. 2020. "When is Discrimination Unfair?" AEA RCT Registry. September 22.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez and Stefanie Stantcheva. 2015 “How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments.” *American Economic Review* 105(4): 1478-1508.

Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association* 11(3): 495– 524.

Lefgren, Lars J., David Sims and Olga Stoddard. 2016. "Effort, luck, and voting for redistribution". *Journal of Public Economics* 143: 89-97.

Hedegaard, Morten Størling and Jean-Robert Tyran. 2018. "The Price of Prejudice" *American Economic Journal: Applied Economics* 10(1): 40–63.

Lippens, Louis, Stijn Baert, and Eva Derous. 2021. "Loss Aversion in Taste-Based Employee Discrimination: Evidence from a Choice Experiment." *IZA discussion paper* no. 14438.

Mas, Alexandre. 2017. "Does Transparency Lead to Pay Compression?" *Journal of Political Economy* 125(5): 1683-1721.

Oprea, Ryan and Sevgi Yuksel 2021. "Social Exchange of Motivated Beliefs" *Journal of the European Economic Association*, forthcoming.

Sauer, Carsten. 2020 "Gender Bias in Justice Evaluations of Earnings: Evidence From Three Survey Experiments" *Frontiers in Sociology* 7(5):22.

Stantcheva, Stefanie. 2021. Understanding Tax Policy: How do People Reason? *Quarterly Journal of Economics* 136(4): 2309–2369.

Tilcsika, András 2021. "Statistical Discrimination and the Rationalization of Stereotypes" *American Sociological Review* 86(1): 93–122.

Figures and Tables

Figure 1: Fairness Ratings by Type of Discrimination and *Justifiability*

p-values:

Less- versus more justifiable treatments:

Overall: p=.000

Within taste-based: p=.000

Within statistical: p=.000

Taste versus Statistical Discrimination:

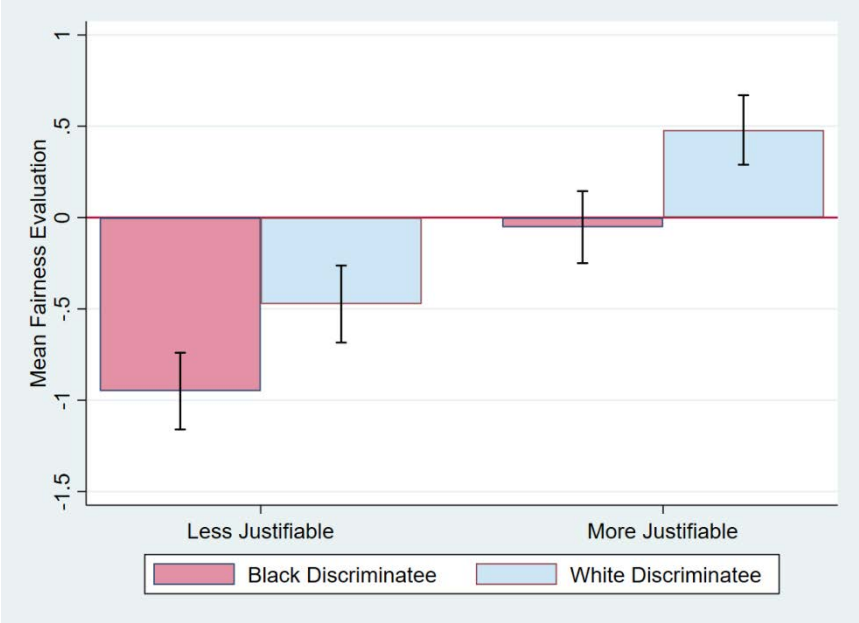
Overall: p=.971

Within Less-Justifiable: p=.779

Within More-Justifiable: p=.710

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent.

Figure 2: Fairness by *Justifiability* and Discriminatee Race



p-values:

Black versus White Treatment:

Overall: p=.000

Within Less-Justifiable: p=.002

Within More-Justifiable: p=.000

Less versus More Justifiable Treatment:

Overall: p= .000

Within Black Discriminatees: p=.000

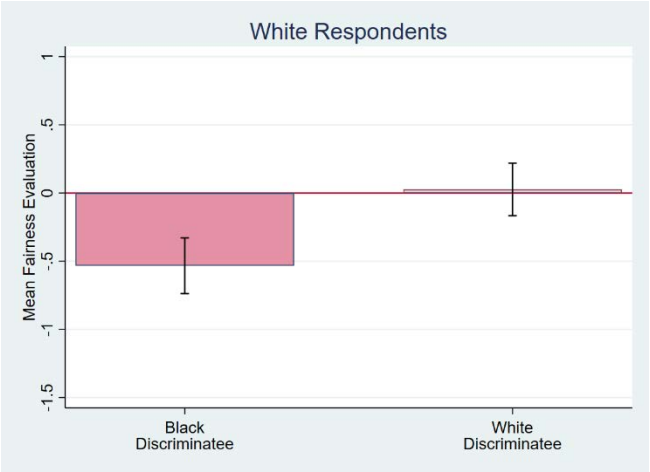
Within White Discriminatees: p=.000

Notes : This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent.

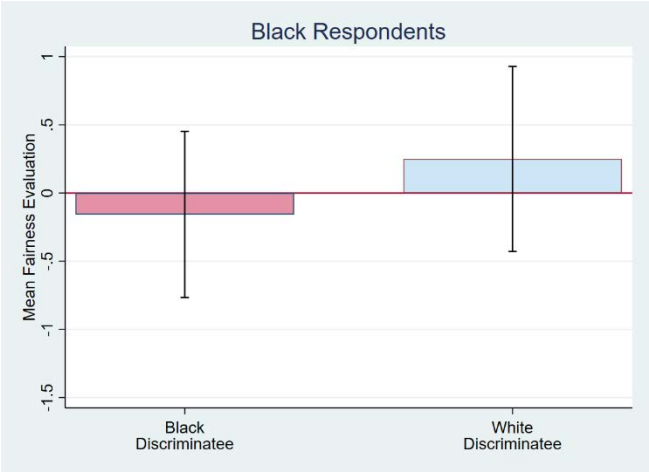
Within Black Discriminatees, less-justifiable scenarios are 0.898 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.953 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields p = .679.

Figure 3: Fairness Ratings by Respondent Race and Discriminatee Race

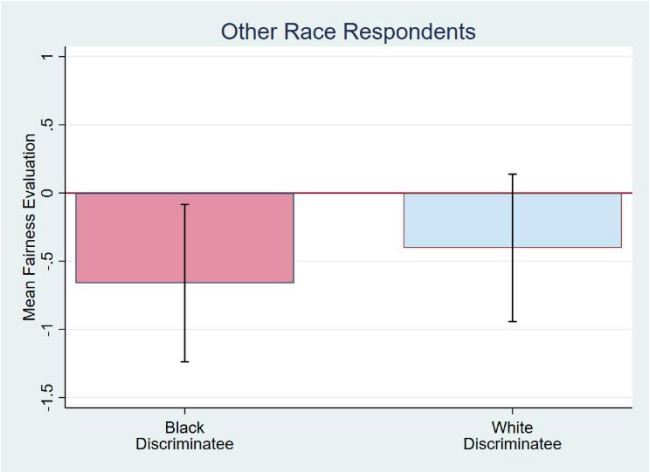
(a)



(b)



(c)



p-values:

Black versus White Treatment:

Overall: Overall: p=.000

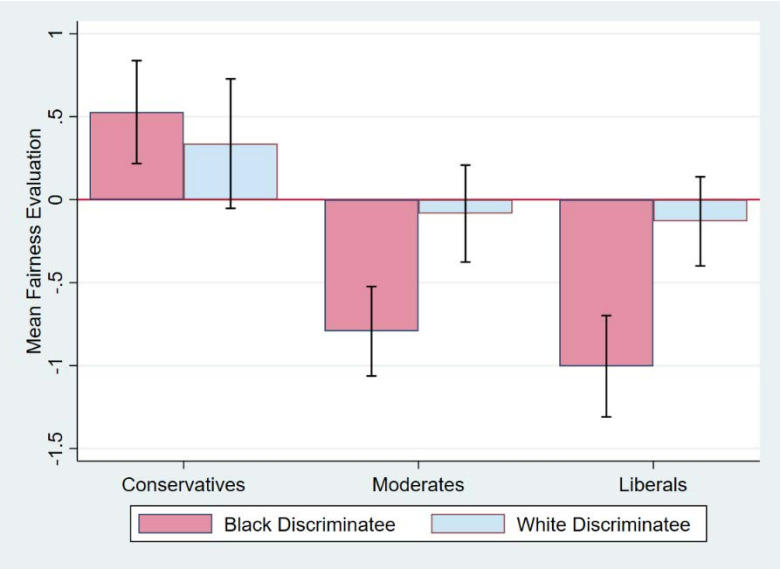
Within White respondents: p=.000

Within Black respondents: p=.365

Within Other respondents: p=.506

Note: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) across all three racial groups yields p = .739.

Figure 4: Fairness Ratings by Political Orientation and Discriminatee Race



p-values:

Black versus White Treatment:

Overall: *p*=.000

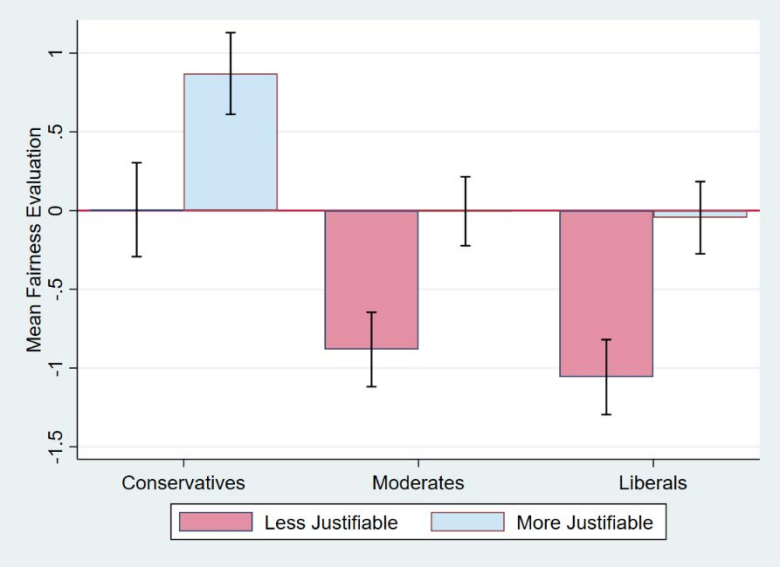
Within Conservatives: *p*=.448

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the discriminatee race effect (i.e., the Black treatment) between moderate and liberal respondents yields $p = .567$. A test for equality between conservatives and (moderates + liberals) yields $p = .001$.

Figure 5: Mean Fairness Evaluations of Less- versus More-Justifiable Discrimination Scenarios, by Respondent's Political Leaning



p-values:

Less versus More Justifiable Treatment:

Overall: *p*=.000

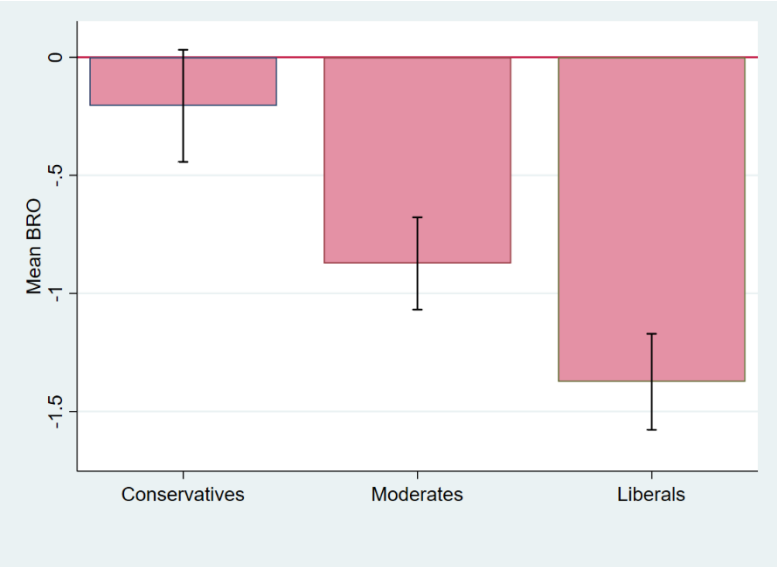
Within Conservatives: *p*=.000

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Notes: This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of the Less versus More *Justifiability* Gap across Conservatives, Moderates, and Liberals yields $p = .590$.

Figure 6: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning



p-values:

Conservatives vs. Moderates: *p*=.000

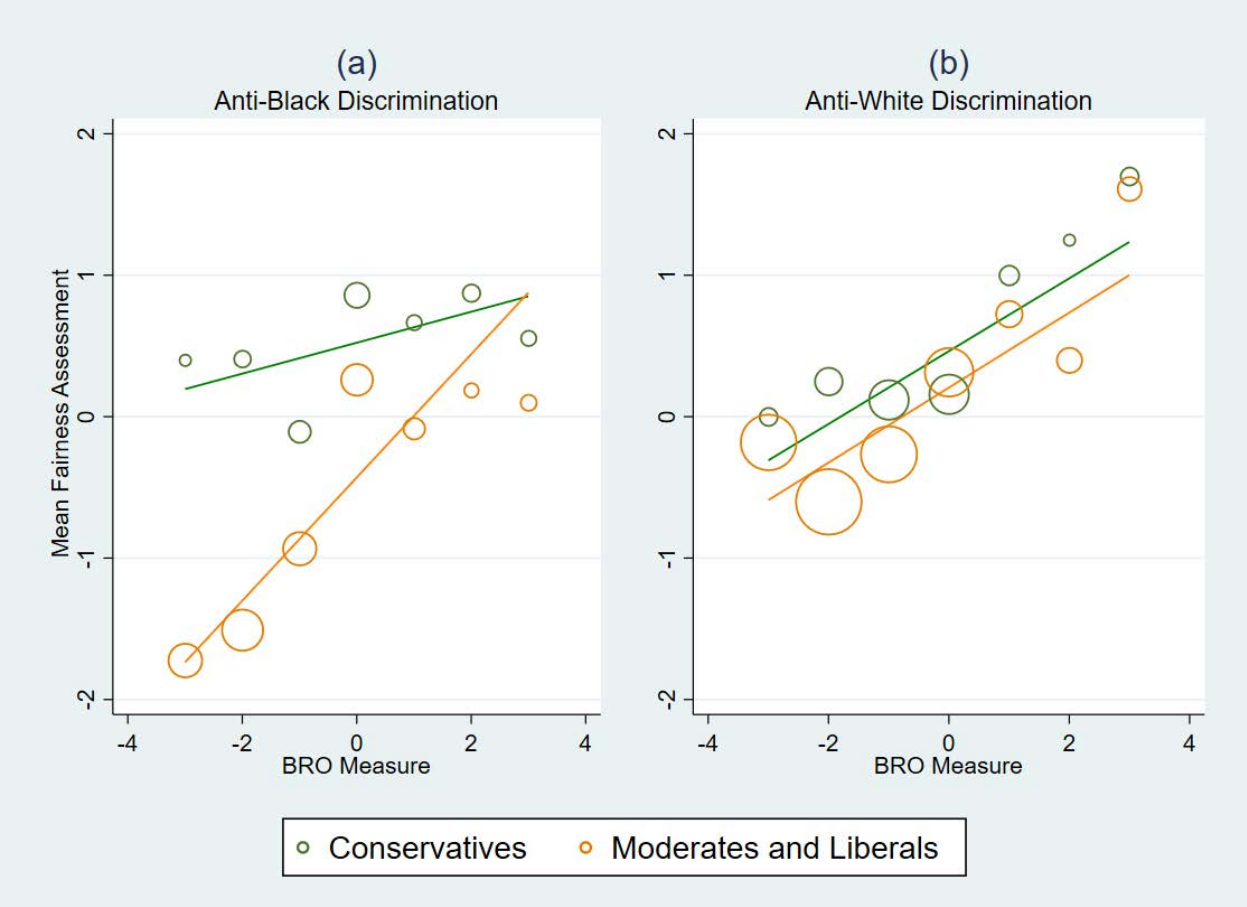
Moderates vs. Liberals: *p*=.001

Liberals vs. Conservatives *p*=.000

Notes:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). This figure is based on only Stage 1 observations. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .577$.

Figure 7: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race



Notes: Symbol size is proportional to the number of respondents. Sample is restricted to Stage 1 fairness assessments only. The *p*-values below are clustered by respondent, except for those pertaining to Panel (c).

- Panel (a), Discrimination against Black Applicants
  - For Conservatives: slope = 0.109, $p = .218$
  - For Moderates and Liberals, slope = 0.436, $p = .000$
- Panel (b), Discrimination against White Applicants
  - For Conservatives: slope = 0.257, $p = .094$
  - For Moderates and Liberals, slope = 0.265, $p = .000$

Political leaning subsamples for anti-Black discrimination:

Conservatives vs. Mods-Libs, BRO = -3 only ($p = .000$)

Conservatives vs. Mods-Libs, BRO = -2 only ($p = .000$)

Conservatives vs. Mods-Libs, BRO = -1 only ($p = .658$)

Conservatives vs. Mods-Libs, BRO = 0 to +3 combined, only ($p = .000$)

Table 1: Estimating the Combined Effects of Utilitarian and Rules-Based Determinants of Fairness

	All	All	Conservatives	Conservatives	Moderates and Liberals	Moderates and Liberals
	(1)	(2)	(3)	(4)	(5)	(6)
Less justifiable	-0.925***	-0.953***	-0.865***	-0.800***	-0.947***	-1.004***
	(0.0665)	(0.0968)	(0.139)	(0.190)	(0.0757)	(0.112)
Black discriminatee	-0.505***	-0.532***	0.190	0.251	-0.790***	-0.848***
	(0.129)	(0.139)	(0.250)	(0.266)	(0.143)	(0.157)
Less justifiable × Black discriminatee		0.0551		-0.122		0.116
		(0.133)		(0.276)		(0.151)
Constant	0.466***	0.480***	0.770***	0.738***	0.365***	0.393***
	(0.0931)	(0.0968)	(0.202)	(0.205)	(0.104)	(0.109)
Observations	1,284	1,284	340	340	944	944
R-squared	0.076	0.076	0.055	0.055	0.110	0.110

Notes: Regression results are based on Stage 1 fairness assessments only. One star indicates a ten percent significance level, two stars indicate a five percent level, and three stars indicate a one percent level. Standard errors are clustered by respondent.