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, Revised February 2023

The authors thank Catherine Weinberger, participants in the 2022 Trans-Pacific Labor Seminar and seminar participants at Drexel University, Princeton University, and the University of Georgia 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, Revised February 2023
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 react in very similar ways to the employer's motivations for the action, 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 pre-registered models of fairness –(simple) utilitarianism, race-blind rules (RBRs), racial in-group bias, and belief-based utilitarianism (BBU) – and show that the latter two are inconsistent with major aggregate patterns in our data. Instead, we argue that a two-group framework, in which one group (mostly self-described conservatives) values employers' decision rights and the remaining respondents value utilitarian concerns, explains our main findings well. In this model, both groups also value applying a consistent set of fairness rules in a race-blind manner.

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
University of California, Santa Barbara
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 (Bertrand and Duflo, 2017). 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 behavior and support for public policies (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 much less attention.

To study this question, 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 consistency of four pre-registered models of perceived fairness with the patterns we observe. Finally, we provide a simple, non-preregistered, two-group interpretive framework that provides a convenient summary of all our empirical results.

Our main findings are as follows. First, subjects' self-identified political leanings have large effects on their overall acceptance of discriminatory actions, with conservatives being much more accepting of the discriminatory actions we depict than moderates and liberals. Second, regardless of their political leanings, our respondents care about the detailed motivations behind a discriminatory action (holding the act's consequences constant). Specifically, while the presence of taste-related versus statistical factors does not reliably predict subjects' fairness assessments, other aspects of the discriminator's motivations have robust and sizable effects. For example, 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. Similarly, 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 very similar across all political groups, and the effects do not depend on the race of the discriminatee.

Third, our moderate and liberal respondents exhibit a strong *discriminatee race effect*: they disapprove more of anti-Black than anti-White discrimination. This effect is absent among conservatives, who rate the discriminatory acts we depict as slightly more fair than unfair, regardless of the discriminatee's race. Fourth, among the four models of perceived fairness we evaluate –(simple) 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. Fifth, all three political groups in our sample –liberals, moderates, and conservatives—exhibit a strong desire to apply race-blind rules when comparing the fairness of different discriminatory actions. Only liberals and moderates, however, exhibit utilitarian preferences (which assign higher fairness ratings to actions that shift income from high- to low-income groups).

Sixth, to make sense of all these findings, we propose an interpretive framework with two equally sized groups of respondents who collectively care about three fairness criteria: race-blind procedural fairness (RBRs), utilitarianism, and a (non-preregistered) ethic that values employers' decision rights. Both of our groups strongly value race-blind procedural fairness. In addition to this, Group 1, or *Business Rights Advocates*, also care about employer decision rights, but place no value on utilitarian objectives. Group 2, or *Utilitarians*, value utilitarian objectives but exhibit no detectable support for employers' decision rights. Group 1 are predominantly (but not exclusively) self-identified conservatives, while Group 2 is a large subset of moderates and liberals.

Finally, we notice that – unlike Group1 – Group 2 subjects (who are all moderate or liberal) could face a conflict between the two fairness criteria (utilitarianism and race-blind rules) they care about: Objecting more strongly to anti-Black than anti-White discrimination for utilitarian reasons may not feel race-blind. To assess how Group 2 makes this tradeoff, we leverage the fact that a large share of our

subjects experiences a switch in the race of the person being discriminated against during the experiment. Under the assumption that subjects only become aware of their desire to make race-blind fairness assessments after they are exposed to a discriminatee of a second race, we are able to use random assignment of our *race* treatments to estimate that Group 2's fairness assessments place roughly equal weight on these two criteria when they conflict.

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, Schildberg-Hörisch et al. 2022). Some of these authors have argued, for example, that effort- and 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 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 the fairness of income differentials; these

[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.

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 race-based policies like affirmative action; these beliefs also account for much of the partisan divide in policy support. Informational treatments designed to change people's beliefs, however, have limited effects on policy support. Our paper differs from these three 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 public policies. Second, we consider a broader set of implicit fairness models that people might use to assess either actions or policies. Specifically, we 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 set of rules when mapping intentions and actions into fairness levels – play important roles in non-consequentialist ethics such as those studied by Andreoni et al. (2019). In our 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 (and potentially which types of race-relevant public policies) are perceived as fair or unfair.[3]

[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.

[3] 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.

Finally, our analysis relates to ongoing debates among both economists and legal scholars about which forms of discrimination are more 'egregious' than others (and therefore perhaps more deserving of policy remedies or legal sanctions.) For example, in a recent review article, Bertrand and Duflo (2017) provide the following description of a common view among economists:

> While taste-based discrimination is clearly inefficient…, statistical discrimination is theoretically efficient and, hence, more easily defendable in ethical terms under the utilitarian argument. Moreover, statistical discrimination can also be argued to be "fair" in that it treats identical people with the same expected productivity (even if not with the same actual productivity) [equally] and is not motivated by animus. In fact, many economists would most likely support allowing statistical discrimination as a good policy, even where it is now illegal… (Bertrand and Duflo 2017, p. 312).[4]

In the case of legal debates and proceedings, influential decisions like Griggs v. Duke Power Co. (1971) have maintained that *intent* is not essential for an act or policy with race-based consequences to be unlawful, instead these decisions maintain that disparate *impact* is enough. This disparate impact principle continues to be contested, however.[5] Our paper contributes to both these economic and legal debates by describing how *a broad sample of Americans* perceives the fairness of different types of discriminatory actions. We find that (a) the detailed intentions underlying a discriminatory action *do* matter for peoples' fairness perceptions, but that (b) whether the action was motivated by someone's racial animus ('tastes') is not, on its own, a reliable guide to an action's perceived fairness.

Section 2 of the paper describes our survey design, data collection, and sample characteristics. Section 3 presents some basic facts about fairness perceptions: How do perceptions vary with respondent characteristics, survey treatments, and interactions between the two? Section 4 describes four simple, preregistered models of fairness and compares their implications to subjects' aggregate response

[4] The word "equally" is not present in Bertrand and Duflo's text; we have inserted it to convey what we believe is their meaning.

[5] Despite these disputes, there seems to be wide agreement that the presence of racial or other animus would make the same discriminatory act *more* egregious.

patterns. It shows, --among other things—that our Group 2 respondents (*Utilitarian*s) care about two criteria –utilitarianism and race-blindness-- that sometimes conflict. Section 5 then uses within-subject variation in the discriminatee's race to estimate the relative weight these *Utilitarian* subjects place on the two criteria they care about. Section 6 concludes.

2. Design and Implementation

2.1 Survey Structure

Before starting our survey, all our subjects were 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).

Next, our subjects are randomly assigned to read four vignettes in which an employer, "Michael" (or "Andrew") makes a hiring decision between a White and a Black applicant.[6] These scenarios are designed to represent canonical examples of taste-based and statistical discrimination. We define taste-based discrimination as a decision that is based on *someone's* racial animus, and distinguish two forms: *Less-justifiable* taste-based discrimination is based on the *employer's own distaste* for people of a particular race. *More-justifiable* taste-based discrimination occurs when the employer accommodates his *customers'* distastes for a particular race.[7] Statistical discrimination, on the other hand, is based on differences in expected job performance. *Less-justifiable* statistical discrimination is based on low-quality information about the relative performance of two racial groups; we frame this as a non-

[6] Michael and Andrew appear to be the most common male names that are relatively race-neutral. Between 2011 and 2016, they ranked in the top 2-6 names for White men and the top 6-12 names for Black men in New York City birth names.

[7] Notice that –to the extent that it is costly to attract new customers-- our employer's decision is profit-maximizing in the more-justifiable version of taste-based discrimination, but not in the less-justifiable version.

quantitative statement from a single, non-expert source (a 'neighbor') about problems *others* experienced when employing White or Black employees, such as lateness and lack of attention to detail. *More-justifiable* statistical discrimination is based on "reliable statistics" from "a large and experienced network of local business owners" who frequently hire for the same type of opening as in the scenario.[8] In both cases, the 'justifiability' rankings of the more detailed reasons for the action are based on our own priors regarding how respondents would react. After reading each vignette, the respondent is asked to rate the fairness of the employer's hiring decision on a scale from 1 to 7. As in Alesina et al. (2020, 2021), these questions have no material consequences.[9]

The four scenarios encountered by each respondent are presented in two Stages. In Stage 1, subjects were assigned with equal probability to one of the four possible treatment combinations: SW, TW, SB, and TB, where S and T represent statistical and taste-based discrimination, and W and B indicate the race of the discriminatee. Within Stage 1, the subjects encounter the less- and more-justifiable versions of discrimination in random order. In Stage 2, subjects were randomly assigned to one of the three treatment combinations they did not encounter in Stage 1, and again encountered the more- versus less-justifiable forms in random order.[10] Thus, as illustrated in Appendix A1.2, two thirds of the respondents encountered a switch in the discriminatee's race, and two thirds encountered a switch between taste-based and statistical discrimination.

Our survey concludes with Stage 3, which asks all subjects the same questions. First, in an open-text question, we remind respondents of the final scenario they encountered and invite them to explain the

[8] As framed in our vignettes, low-quality information could be interpreted by our subjects as an unbiased signal with high variance, as a biased signal (Bohren et al. 2019), or even as a signal whose bias is motivated by someone's racial animus. In all cases we would expect that relying on such signals will be seen as less fair than using the information described in our high-quality statistical scenario.

[9] Cappelen et al. (2019) and Almås et al. (2020) create real *income* inequality (for example, between two MTurk workers), then give third-party subjects options to reduce this inequality. Applying this methodology to discriminatory incidents would raise serious ethical concerns. It is also unclear how we could manipulate the *motives* of a real discriminator, which matter a lot for peoples' fairness assessments.

[10] To make the scenarios more realistic, the name of the employer also switches between the Stages. Specifically, half the employers are Michael and the other half Andrew in Stage 1; this assignment is random. In Stage 2, the name of the employer switches to the other, unused name for all respondents.

fairness assessment they made. Next, we use the following question to elicit subjects' assessment of Black people's relative economic opportunities (BRO):

> Please consider the following question without referring to any of the previous survey items, and then select the rating that best corresponds to your answer:
>
> *All in all, in the United States, how would you compare the economic opportunities available to Black and White people? Black people have:*
>
> *Much less / Less / A Little Less / Roughly equal / A little more / More / Much More opportunity than White people.*

Finally, we collected information on the subjects' age, education, race, gender, and political affiliation.

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:

> 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".

As noted, 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 different from Stage 1.[11] White scenarios were identical to Black scenarios except that the races of the discriminator and the discriminatee are reversed. As noted, *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). The exact wording of these and all our scenarios is provided in Appendix 1.1.

[11] Exposing subjects to four scenarios (rather than one) has three main benefits. First, it gives us more fairness assessments, while preserving the option to use only each subject's first treatment for pure cross-subject comparisons. Second, as illustrated in Section 5.1, it allows us to test for a specific but plausible form of experimenter demand effects. Third, as illustrated in Section 5.3, it allows us to assess the relative importance subjects assign to utilitarian, versus race-blind fairness criteria when those criteria conflict.

2.3 Implementation and Representativeness

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. 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.[12] 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. Finally, respondents were exposed to each vignette for at least 30 seconds before being allowed to submit their fairness assessment. In all, we received 779 responses; during data cleaning we dropped 137 of these, leaving us with a final count of 642 responses in our analysis sample.[13]

In Appendix 2.2, we present summary statistics of our survey respondents and compare them to adults in the 2019 American Community Survey (ACS) and the 2020 General Social Survey. Compared to the ACS, our sample of MTurkers is quite regionally representative, a little more male, and a little more likely to be either White or Black. Our respondents are also considerably better educated and much more likely to be between 25 and 44 years of age than U.S. adults in general. For the most part, these are well known features of the MTurk population.[14] Comparing our subjects' political orientations to the GSS is more difficult because –despite the similarity of the survey questions—the middle category differs

[12] 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.

[13] The main reasons for excluding responses were (i) a pinged location suggesting that the respondent was not U.S. based, and (ii) indications that the response was automated (for example, the IP address attempted our survey more than once, or the response copied and pasted word-for-word sentences from the vignettes into their open text answer.) Summary statistics on the final sample's characteristics (race, gender, education, political orientation, and location within the U.S.) can be found in Appendix 2.

[14] For additional discussions of the representativeness of MTurk samples, see Kuziemko et al. (2015), Arechar et al. (2017), and Everett et al. (2021).

between the two surveys: "moderate" in our case versus "moderate, middle of the road" in the GSS.[15] Ignoring this difference in phrasing, it would appear that our MTurk respondents are politically more 'extreme' than GSS respondents, with more candidates selecting the two extreme categories and far fewer selecting the middle one. However, GSS respondents could be attracted to the attracted to the 'middle of the road' label.[16] In sum, our MTurk-based sample differs from the U.S. population in substantial and mostly well-known ways. In Appendices 7 and 8 we estimate the implications of these differences by re-weighting our main results to match the American Community Survey and General Social Survey respectively. The results are very similar.

2.4 Question Order Effects

In all multi-part surveys, but especially in contexts like ours where framing and experimenter demand effects might play a large role, the order in which respondents encounter different questions could have large effects on the respondents' answers. We address this issue in detail in Appendix 3, which shows 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.[17] 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 on subsequent scenarios. 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

[15] Since the ACS does not collect information on political opinions or affiliations, we are forced to use the GSS (with its much smaller sample size) to assess the political representativeness of our population. Our political party preference question is not comparable to the GSS's, but (with the exception of this middle category) our political leaning question is identical to the GSS's (see Table A2.2 for details).

[16] In addition to this possible difference in variance, there is also some suggestion that, on average, MTurkers are somewhat more liberal than GSS respondents. Almost identical shares of MTurkers and GSS respondents choose some degree of conservative leaning (ranging from slight to extreme), but many more MTurkers choose some liberal leaning (47.3 versus 30.2 percent). This difference could reflect the relative youth of MTurkers.

[17] 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.

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.[18]

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' Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1. In the next two Sections of the paper (3 and 4), we will eliminate the influence of these 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 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.[19] To account for within-subject correlation of responses, all standard errors are clustered by subject.

[18] 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.

[19] 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

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

In Figure 1, we show how the mean perceived fairness of discriminatory acts varies with respondents' characteristics. To maximize our sample size for these initial comparisons, we pool responses across all four treatment combinations (SW, TW, SB, and TB) as well as the more- and less-justifiable versions of both types of discrimination. In short, Figure 1 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; we explore this correlation in Appendix 4 and show that higher levels of education mostly reflect a higher 'set point' for all fairness assessments in the following senses. First, regardless of political leaning, more-educated individuals rate *all* the scenarios they encounter as more fair than less-educated individuals. Furthermore, in contrast to political leaning –which has strong effects on how our subjects respond to some of our treatments-- highly-educated subjects respond to all our treatments in very similar ways to less-educated subjects.[20]

Finally, Figure 1 shows that respondents' political leaning is strongly related to 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).[21] Mean fairness assessments across U.S. political *party* preferences (e.g., Democrats versus Republicans) exhibit similar patterns.[22] Since the Independent group could include people with both extreme right- and left-wing

---

the four scenarios each person encountered, although the standards errors are somewhat higher. See for example Figure A5.1.1, which replicates Figure 1 using first-scenario data only.

[20] For example, Appendix 4 shows that more-educated respondents' greater tolerance of discriminatory acts is not confined to discrimination against a particular race—it applies equally to anti-Black and anti-White discrimination. We also show 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.

[21] To account for the fact that our data contain multiple observations per respondent, all p-values in the paper are clustered by respondent.

[22] There is a statistically insignificant non-monotonicity with respect to party preference, with Independents being more opposed to discrimination than Democrats.

orientations, we use conservative-liberal leaning rather than party affiliation to categorize respondents' political preferences in the remainder of the paper.[23]

3.2 Treatment Effects

In Figure 2, 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.[24] 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."[25] 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 2, 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 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.[26] In contrast, the less

[23] Of the respondents who identify themselves as "Independent" within our sample, about 8.43% suggest they are either "extremely liberal" or "extremely conservative." Furthermore, all the results by political party are very similar, with occasional non-monotonicities similar to Figure 1(e), where Independents appear to be to the left of Democrats.

[24] Figure 3 shows that the effects of *justifiability* are virtually identical for White versus Black discriminatees.

[25] As noted, the subjects saw these verbal descriptions, associated with the numerals 1 through 7.

[26] 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.

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 3 we turn our attention to the *race* treatment—i.e. the race of the person who was discriminated against. Motivated by Figure 2 (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 3 shows that respondents view the same discriminatory acts more negatively when they are directed at Black than at White job applicants. We call this phenomenon the *Discriminatee Race Effect* (DRE). The DRE shown in Figure 3 is substantial in magnitude, amounting to about 0.5 fairness units or 0.263 standard deviations, and highly statistically significant (p= .002 and .000 within the *less* versus *more* justifiable forms of discrimination, respectively).

3.3 Heterogeneity: Discriminatee Race Effects by Respondent Race and Political Orientation

While the effects of the *race* treatment shown in Figure 3 are interesting, these effects may not be the same for all types of respondents. For example, Black respondents might react more negatively than White respondents to discrimination against Black job applicants. To explore this issue, Figure 4 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 however suggest that both groups respond to the race of the discriminatee in much the same way as White respondents do.[27] In sum, Figure 4 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

[27] Interestingly, Figure 4 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 imprecise.

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 5 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.[28] A final striking finding from Figure 5 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 exploit this fact to simplify our analysis by comparing just two political groups—conservatives versus moderates/liberals.

4. 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 show that two of the models –racial in-group bias and BBU-- are inconsistent with some key patterns in our data. After examining subjects' open-text responses for clues that might explain these inconsistencies, we then propose a two-group framework with three fairness criteria that does a better job of accounting for the facts we have documented. 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.

[28] The confidence interval for anti-Black discrimination is bounded above zero; the mean fairness assigned to anti-White discrimination is almost identical, but not quite significantly different from zero.

4.1 (Simple) Utilitarianism

Since utilitarian models belong to the consequentialist family of ethical systems, fairness in these models depends on outcomes of actions, not on intentions or justifications. Since our Tastes vs. Statistical and less-versus-more justifiable treatments refer to the *reasons* for the employer's actions, and since the consequences of these actions –i.e. *which* worker got the job—are statistically balanced across all our observations, the fairness ratings of purely utilitarian respondents should not differ between any of the Taste-based, Statistical, or less- and more-justifiable forms of discrimination. A second property of utilitarian fairness models is that they use a *social welfare function* to map consequences into fairness levels. In the simple utilitarian model we consider here, this social welfare function is a strictly concave function of the incomes of the people depicted in our scenarios. Since mean racial income differences indisputably favor White people, utilitarian respondents should be more tolerant of acts of anti-White than anti-Black discrimination. Importantly, this prediction holds even if we account for the direct effects of discrimination on employers' utility: Taste-based discrimination may raise the utility of the employer, and statistical discrimination may raise profits. This because the employer in our scenarios is always White when the discriminatee is Black, and *vice versa*.

We refer to the type of utilitarianism described in this subsection as 'simple' because it is based purely on racial *income* differences, which are publicly verifiable information. Real respondents might, however, base their ideas of deservingness on criteria other than income (such as *opportunities*)—and real respondents could also have inaccurate and widely varying beliefs about racial gaps in both income and opportunity (Davidai and Walker 2021, Kraus et al. 2017, 2019). We will consider these possibilities under the heading of *belief-based utilitarianism* (BBU) below.

Turning to the evidence on simple utilitarianism, Figure 3 has already shown that respondents in general *do* view discrimination against Black applicants more harshly than discrimination against White applicants. That said, Figure 5 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 our (simple) utilitarian model is consistent with moderates' and liberals' fairness assessments, but is

not consistent with conservatives' fairness statements.[29] Utilitarianism also cannot account for the large justifiability effects on perceived fairness that are documented in Figures 2 and 3.

4.2 Racial in-group bias

The phenomenon of in-group favoritism, where people value actions that benefit members of their own identity group more than actions benefiting others, has been extensively documented (Luttmer 2001, Chen and Li 2009, Fong and Luttmer 2009, 2011, Everett et al. 2015). While a variety of models could explain this behavior, a simple one, based on social preferences, modifies the preceding utilitarian model in a straightforward way: 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. In our experiment, respondents who exhibit racial in-group bias should view the discriminatory acts we depict as less fair when the fictitious discriminatee shares the respondent's race.[30]

As Figure 4 has already shown, we do not have the statistical power to test these predictions for the respondents in our Black or Other racial categories.[31] 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. Interestingly, when we focus our attention on the subset of White respondents who identity as conservative, this strong rejection of in-group bias no longer holds: As shown in Figure A5.2.1, White conservatives rate discrimination against Black people as 0.405 units *more* fair than discrimination against White people.

---

[29] An insignificant discriminatee race effect for conservatives could imply that that they are not utilitarians at all (i.e. they do not use a social welfare function (SWF) to make fairness assessments). Alternatively, conservatives could be utilitarians with a linear SWF. A final possibility, considered under *belief-based utilitarianism* (BBU) below, is that conservatives' SWF depends on something other than income (such as, for example, perceived relative opportunities).

[30] 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.

[31] 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.

This difference is however not statistically significant at conventional levels ($p = .134$). 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 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.[32] Finally, rules-based models of fairness are *race-blind* when the rules that assign fairness to actions and intentions 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. For example, if using low-quality statistical information is x units less fair than using high-quality information, x should be the same regardless of the race of the discriminatee.

[32] 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.

The fairness ratings of our respondents are consistent with the use of race-blind rules (RBRs) in at least three ways.[33] First, the effects of our *justifiability* treatments in Figure 2 strongly support the idea that respondents care about the employer's motivation for discriminating against a job applicant. Importantly, our experimental design ensures that the *justifiability* effects in Figure 2 hold the consequences of the discriminatory action constant: While the material consequences of not being hired could, for example, vary with the discriminatee's race (because of differences in outside labor market options), notice that Figure 2 varies only the *reasons* for not being hired: discriminatee race is balanced between the motivation and justifiability treatments due to random assignment.

Second and more strikingly, Figure 3 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*: (-0.953 versus -0.898 fairness points for White versus Black discriminatees respectively, with $p = .679$ for a test of equality). 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.[34] A final, remarkable feature of our respondents' apparent adherence to race-blind rules is that it applies just as strongly on both sides of the U.S. political divide. We show this explicitly in Figure 6, 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 *per se* do 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

[33] In Sections 5.2 and 5.3, 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.

[34] 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 $p = .319$.

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, even though 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).[35] 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.[36] From a modeling perspective, BBU 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 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 is consistent with BBU is presented in Figure 7, which draws on the BRO question in Stage 3 of our 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" (minus 3 in Figure 7) to "much more opportunity" (plus 3 in Figure 7). Figure 7 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

[35] 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.

[36] Our survey design does 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.

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.[37]

To assess whether BRO differences can actually account for the partisan gap in fairness assessments, panel (a) of Figure 8 shows respondents' fairness ratings for anti-Black discrimination by BRO categories, separately for conservatives and moderates/liberals.[38] If BRO accounts for the large partisan gap, we should see little or no partisan gap *within* the BRO categories: Instead, the partisan gap should be explained 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 8 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 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). This partisan gap at the bottom of the BRO distribution is highly statistically significant. 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.

[37] 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 contributions relative to these papers are that we study the fairness of individual (discriminatory) actions (not public policies), we demonstrate the key role of the intentions behind an action in determining its perceived fairness, and we test the BBU model that underlies the idea that changing beliefs about opportunities can change support for policies.

[38] Starting in this subsection, we combine moderates and liberals into a single group to simplify the presentation and preserve statistical power. Interested readers can view a version of Figure 8 with all political groups in Appendix 5.3; all the qualitative results discussed below are similar for moderates and liberals individually, as well as the combined group.

A third and even more surprising piece of evidence against the "BRO hypothesis" emerges from panel (b) of Figure 8, 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 (a) correlated with their political affiliations and (b) sometimes predictive of their fairness assessments, 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. This rejection of the BRO hypothesis in our data might help explain why interventions designed to change beliefs about relative opportunities do not have robust effects on support for race-based policies, *even when the interventions change beliefs* (Alesina et al. 2021; Haaland and Roth 2021).

4.5 What Motivates Respondents Who Say Discrimination is Fair?

To try to make sense of the unexpected findings in Figure 8, we first observe that a large subset of our respondents (when classified by beliefs and political orientation) exhibit a common pattern of fairness assessments: they assign roughly equal, non-negative fairness to both anti-Black *and* anti-White discriminatory acts. In Figure 8, this group of respondents includes all self-identified conservatives, *plus* the moderates and liberals with BRO ≥ 0. Together, these respondents (henceforth Group 1) represent 48.9 percent of all respondents. In Figure 8 they are indicated by the hollow circle markers.

To understand what types of fairness criteria might account for these respondents' fairness assessments, we then turned to our respondents' open-text explanations of the last scenario they encountered. As described in Appendix 6, we manually classified these explanations into three broad categories, separately for respondents who rated discrimination as "unfair" or "very unfair", versus respondents who said it was "fair" or "very fair". For the latter group, the most common type of response was some variation of '*the business must thrive*', such as:

> "The hiring decision was fair because any individual in Michael's shoes would do anything within their power to protect their business by all means necessary."
>
> "A business wants [to] retain customers and high profits. So give them what they expect. Hiring what people prefer is reasonable."
>
> "Andrew needs to do what is best for his business. If he hired the black worker, he'd lose money and perhaps even go out of business."
>
> "To ensure the success of his business, Michael should do everything possible to do so. Is it racist? I don't think so. Michael is free to hire whoever he wants."

Notably, almost all of these 'business must thrive' answers referred to scenarios in which the employer accommodated his customers' discriminatory tastes (i.e. the 'more-justifiable' version of taste-based discrimination).

Closely related, the second most common type of explanation involved some statement of '*employer rights*', including:

> "It's his company he can hire whoever he choses [sic]. He does not have to give an answer to anyone or share his hiring views. He can choose what is best at any time."
>
> "It's his business. He doesn't need to justify to me any of his hiring practices. … Seriously, he does not need to justify his hiring choices."
>
> "Andrew does run the business so it is within his rights to not hire a black man because he doesn't enjoy interacting with them."

> "The employer should have the right to hire who he is most comfortable with regardless of the reasons."
>
> "I've never had a problem with this as long as every business owner is allowed to do it; I don't feel comfortable in all-black establishments so just I don't go to them. They'd be uncomfortable, I'd be uncomfortable, just let them have their thing. What's the problem?"
>
> "Employers are allowed to hire whoever they see as best for the job."

Notably, these 'employer rights' explanations were expressed with respect to *all* forms of discrimination, including discrimination based on the employer's own tastes.

Taking all the above responses together, we propose that Group 1's high acceptance of discriminatory actions – regardless of the target of the action – is consistent with a fairness system that prioritizes individual decision rights (regardless of those decisions' effects on others), especially for business owners. As shorthand, we therefore label Group 1, defined above, as *Business Rights Advocates*.

4.6 An Interpretive Framework that 'Works'

With this decision-rights-based ethical model in hand, we can now propose a provisional, ex post interpretive framework that ties together all the fairness assessment patterns we have documented in the paper. We hasten to remind readers that –by logical necessity—this framework is not the only one that can account for all those patterns. We offer it primarily as a simple mnemonic device that summarizes the main facts we have established, and as a jumping-off point for future research on these questions.

In our proposed framework, there are two main groups of respondents. Group 1 (the "Business Rights Advocates", accounting for 48.9% of all respondents) includes all conservatives**,** plus moderates and liberals with BRO ≥ 0. These respondents are, on average, accepting of all the discriminatory actions we depict, regardless of the race of the discriminatee. Members of Group 1 justify these fairness assessments as protecting or raising profits and preserving employer rights. Members of Group 1 do not appear to be motivated by any utilitarian concerns (either 'simple' or 'belief-based'). Based on our

findings in Section 4.3, however, they share the widespread support across the political spectrum for race-blind procedural fairness (RBRs).

Group 2 or "Utilitarians" are moderates and liberals with BRO<0, accounting for 51.1% of our respondents. These respondents object to both anti-Black and anti-White discrimination, but object more strongly to anti-Black discrimination. Given their beliefs (BRO<0), Group 1's fairness assessments are consistent with both simple and belief-based utilitarianism. This group exhibits no obvious support for employer decision rights. [39] Like Group 1, Group 2 shares a strong desire for race-blind procedural fairness.

5. Reconciling Conflicting Fairness Criteria: Utilitarianism versus RBRs

In the previous Section, we proposed a two-group interpretive framework in which one group of respondents (Group 1 – *Business Rights Advocates*) cares only about race-blind rules (RBRs) and individual decision rights, while Group 2 (*Utilitarians*) cares only about RBRs and utilitarian welfare criteria. In this framework, there is a clear potential for conflict between Group 2's two main fairness objectives: Rating anti-Black discrimination more harshly than similar acts of anti-White discrimination may not feel race-blind.[40] In this Section, we use the *race* treatment order effects documented in Section 2.4 to estimate the relative weight that members of Group 2 place on these two fairness criteria. Our identifying assumption is that respondents are not aware of their desire to make race-blind fairness assessments until they encounter a discriminatee from a second racial group, i.e. until they encounter a *switch* in the race treatment.

[39] In open-text answers, respondents who object to discrimination frequently say that it is wrong to base hiring decisions on race, statistical information, or tastes. These respondents frequently use words like "racism", "bigoted", "prejudice", "bias" and "stereotype" in their explanations. References to employer decision rights are essentially absent. See Appendix 6 for a detailed analysis of subjects' open-text responses.

[40] Since they are tolerant of both anti-White and anti-Black discrimination, Business Rights Advocates do not experience a similar conflict when the *race* treatment switches. To see this, consider for example a Business Rights advocate who said it was fair for a White business owner to accommodate his customers' anti-Black discriminatory tastes. In this case, race-blindness would be consistent with saying the same action is fair if the races were reversed, which is exactly how Business Rights Advocates behave in our survey.

To accomplish this goal, we proceed in three steps. First, we document that the race treatment order effects described in Section 2.4 are present in Group 2 but absent in Group 1. Second, we use a simple model of reporting behavior, combined with random assignment of the *race* treatment to interpret Group 2's order effects as a compromise between utilitarianism versus RBRs, and to estimate the relative weight Group 2 places on those two criteria when they conflict. Finally, for readers who may be skeptical of our *ex-post* Group 1 - Group 2 dichotomy, we replicate the preceding steps for the categories used in Section 4 of the paper: conservatives versus [moderates + liberals]. The results are very similar.

5.1 *Race* Treatment Order Effects are Absent in Group 1

In Section 2.4 we demonstrated the existence of race treatment order effects in our entire sample of respondents: Their Stage 2 fairness assessments depend on the *race* treatment they encountered in Stage 1. In Appendix 8.1, we show that there are no such order effects for Group 1: Regardless of the discriminatee race they encountered in Stage 1, members of Group 1 view discrimination as a little more fair than unfair (about +0.5 on a scale from -3 to +3) in Stage 2. Respondents from Group 2, on the other hand, exhibit a more pronounced version of the order effects we saw in the sample as a whole. Specifically, Group 2's Stage 2 fairness assessments of anti-Black discrimination are much milder if they encountered a White discriminatee (as compared to a Black discriminatee) in Stage 1 ($p = .009$).[41] Motivated by this fact, we restrict our attention to Group 2 respondents (*Utilitarians*) in the following sub-section, where we interpret Group 2's *race* treatment order effects as a compromise between the values they place on both utilitarianism and race-blindness.

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

As noted above, subjects who value both utilitarianism and race blindness (e.g. members of Group 2) face a conflict when they experience a switch in the *race* treatment. 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), versus respecting their utilitarian desire to

[41] The race treatment a Group 2 subject received in Stage 1 does not have a statistically significant effect on the subject's ratings of anti-White discrimination in Stage 2.

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^1 represent each respondent i's "pure" utilitarian fairness ratings, B_i^* and W_i^*.

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 pure utilitarian rating (B_i^* or W_i^*) for the *new* racial group, or making the same report they assigned to the other racial group in Stage 1 (being race-blind).[42]

Using this notation, in Stage 2 White-to-Black treatment switchers have the option of reporting their pure utilitarian rating of the group they now face in Stage 2 (thereby setting $B_i^2 = B_i^*$), assigning the same rating they assigned (to the other race) in Stage 1 (thereby setting $B_i^2 = W_i^1$), 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 utilitarian preference and $(1-\alpha)$ 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-

[42] 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.

to-W switchers), their sample means $\bar{B}^*$ and $\bar{W}^*$ in any fixed population (such as Group 2) *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,

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

After restricting our sample to Group 2 respondents, Equations (3) and (4) can then be (separately) solved for α, yielding $\alpha = 0.49$ for the White-to-Black switchers and $\alpha = 0.68$ for the Black-to-White switchers.[43] Thus, W-to-B switchers' Stage 2 ratings of anti-Black discrimination place almost equal weight on race-blindness and utilitarianism. B-to-W switchers, on the other hand, act as if they place slightly more weight on utilitarianism than on race-blindness. The 95% percent confidence intervals for α are [0.243,0.800] and [0.405, 1.075] for W-to-B and B-to-W switchers, respectively. Thus, we cannot reject equal weight on both objectives ($\alpha = 0.5$) for either type of switcher. For W-to-B switchers, we can reject both α=0 and α=1, indicating strictly positive weight on both objectives. For B-to-W switchers, we reject α=0 but not α=1.

Summing up, the *race* treatment order effects we observe among our Group 2 (*Utilitarian*) 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 'split the difference' about equally between these two objectives when making their fairness assessments.

[43] See Appendix 8.1 for the details of the calculations reported in this Section.

5.3 Replicating the Analysis by Political Orientation

In Appendix 8.2, we replicate the preceding analysis, splitting the sample by political orientation (conservatives versus moderates/liberals) with very similar results. Specifically, we show that race treatment order effects are absent among conservatives. Among moderates and liberals, they are stronger than in the full sample. Next, restricting attention to moderates and liberals, we use the same method to calculate α, the relative weight this group assigns to utilitarianism versus race blindness, The point estimates are $\alpha = 0.44$ for the White-to-Black switchers, and $\alpha = 0.62$ for the Black-to-White switchers, with confidence intervals [0.155,0.791] and [0.348, 1.033]. [44] As for Group 2, we conclude that moderates and liberals assign roughly equal weight to utilitarianism versus race-blindness when forced to choose between these two fairness criteria.

6. Robustness

One potential concern about the external validity of our results is the fact that our data were collected in September and October 2020, following a summer of civil unrest related to the murder of George Floyd on May 25, 2020. Together, these events led to a mainstream conversation 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 unusual ways. To check for this possibility, Figure A2.1 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 have been a significant issue for our respondents.

One striking result of our analysis is the large magnitude, statistical significance, and stability of the *justifiability* treatments, documented in Section 4.3: Respondents of all political orientations penalized the less- justifiable forms of discrimination (relative to more-justifiable forms) by the same amount, irrespective of the discriminatee's race. A possible concern with this result is the fact that the

[44] See Appendix 8.2 for the details underlying these calculations.

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 how they rank the fairness of these two types of scenarios, with respect to each other. To address this issue, Appendix A5.1 replicates Tables 1 and 2 using only the first individual 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 8 illustrated some strong and unexpected relationships among subjects' beliefs about relative opportunities (BROs), subjects' political orientation, the race of the discriminatee, and subjects' fairness assessments. Together, these relationships were starkly inconsistent with the belief-based utilitarian model of fairness. To probe the robustness of these results to the fact that Figure 8 combines moderates and liberals into a single group, Figure A5.3 replicates Figure 8, showing separate results for moderates versus liberals. Consistent with other results in the paper, these two groups exhibit very similar response patterns, both of them differing substantially from conservatives' patterns.

In Section 5 of the paper, we used *race* treatment order effects to estimate the relative weight a subset of our subjects assign to utilitarian and race-blind fairness criteria. These order effects could, however, be caused by a particular form of experimenter demand effects that seems quite plausible in our context. To see this, consider the following possibility: 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. Together, these patterns could account for exactly the type of race treatment order effects we observe, where (for

example) respondents' Stage 2 opposition to anti-Black discrimination is reduced if they encountered anti-White discrimination in Stage 1.

In Appendix 7, we provide highly suggestive evidence against this possibility based on the idea 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' 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 shade their own reported political leanings in a more conservative direction on our seven-point scale.

In Appendix 7, 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 *race* treatment order effects we observe.[45] While not conclusive, the stability of subjects' Stage 3 responses to previously-encountered race treatments also suggests the experimenter demand effects are likely not responsible for the strong and robust *justifiability* effects we estimate.

A broader concern that could affect the validity of all our results is the fact that our sample of MTurk respondents was not representative of adult Americans on a number of key dimensions, including

[45] Additional evidence against this demand-effects hypothesis is the fact, documented in Appendix 8.2, that race treatment order effects are absent among conservatives. For experimenter demand effects to explain our results in Section 5, 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 contrast, the fairness reporting model in Section 5 has a 'built in' explanation for conservatives' lack of order effects: Conservatives do not value utilitarian objectives, so they experience no conflict between utilitarianism and the fairness criteria they care about.

age and education (see Appendix 2 for details). While our small sample size limits what we can do to address this issue, Appendices 9 and 10 replicate all our main results (Figures 2-8) two different ways. First, Appendix 9 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 8(a) becomes somewhat stronger and statistically significant. Similar to Figure 8, however, the slope for conservatives remains much lower than the slope for moderates/liberals. Second, Appendix 10 replicates Figures 2-8 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.[46] Despite significant differences in the political mix of the two surveys, all the main results are replicated.[47]

As documented in Appendix P, one of the main differences between the results reported in our paper and the methods proposed in our pre-analysis plan (PAP) involves how our main outcome variable (fairness) is measured. While the PAP proposed a standardized measure of fairness, we realized that a standardized fairness measure centered at the mean fairness level across all respondents and treatments would obscure meaningful cardinal information, relative to the measure we decided to use. Our measure is centered at a conceptually meaningful level --"neither fair nor unfair"-- and our measure's integer values correspond directly to the seven fairness categories respondents could choose from. We made a similar choice with respect to our measure of Black people's relative opportunities (BRO), centering it at "roughly equal opportunities" rather than the sample mean. To see if these decisions had any effects on

[46] In our GSS re-weighting exercise we ignored the difference in the wording of the middle political leaning category between the GSS and our survey. As noted, this might exaggerate the difference between the two surveys, so it should give us an upper bound on the effects of re-weighting. 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 extreme and imprecise weights. The ACS does not ask questions about political orientation or party preference.

[47] The one exception noted with the ACS weights in Appendix 9 does not occur here.

the main relationships established in the paper, Appendix 11 replicates Figures 2-8 using standardized fairness and BRO measures. There is no meaningful difference.

A common critique of non-incentivized survey experiments like ours is that subjects have little incentive to answer the questions thoughtfully, leading to results that are noisy, or simply different from what the same person might offer if they took more time to think about the question. We took a number of precautions to prevent this (see Section 2.3), but it remains possible that many of our respondents gave careless answers that differ from what a thoughtful person would choose. To assess this possibility, Appendix 12 replicates all our main results (Figures 2-8) for 'thoughtful' respondents only, where 'thoughtful' is defined as taking more than the median amount of time to complete the survey. No meaningful differences were evident.

A final important concern --which affects 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 concern, 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. Briefly, Appendices P1-P3 together comprise a "populated PAP" which reports the results of the exact tests specified in the PAP. Appendix P4 summarizes the relationship between the PAP and the paper. In a little more detail, Appendix P4 shows 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 Sections 4.1-4.4 and the main tests thereof (the models' names have changed 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. These are all relatively minor, and the populated PAP results in Appendices P1-P3 strongly suggest they do not matter. Finally, Appendix P4 notes that there are only two PAP hypothesis tests that we decided *not* to include in the main paper and discusses our motivations for those decisions.

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 find, first of all, that conservative respondents are more accepting of discriminatory actions than moderate and liberals. Second, while distinguishing between statistical and taste-based discrimination has been of considerable interest to economists, whether discrimination is motivated by (someone's) tastes or by statistical reasons is not a reliable predictor of assessed fairness. Third and in contrast, respondents of all political leanings *do* care about other aspects of the motivation behind a discriminatory act. Specifically, our respondents agree that acting on one's *own* tastes is less fair than accommodating others' tastes, and that using imprecise or inaccurate statistical information is less fair than precise information. Indeed, respondents of all political leanings penalize these less-justifiable actions by the same amount, and do so regardless of discriminatee race, suggesting a broad area of common ground in how Americans react to different discriminatory actions. Fourth, another important partisan difference is that only moderates and liberals consider the race of the discriminatee when assessing the fairness of a discriminatory act.

Comparing the preceding findings with four pre-registered models of how respondents might make fairness assessments, we find that two of those models – in-group bias models and belief-based utilitarianism –conflict with several key patterns in our data. Using open-text data to identify an unanticipated rationale underlying some subjects' fairness assessments, we propose an *ex post* interpretive framework with two equally-sized political groups and three models of fairness – simple utilitarianism, race-blind rules (RBRs), and employer decision rights – that can account for most of the fairness patterns we observe. In this model, both political groups value using a set of race-blind rules to compare the fairness of different types of discriminatory actions. One group, who we call "Business Rights Advocates" and are mostly conservative, also value employer decision rights. The other group, "Utilitarians" is a large subset of moderates and liberals. They value utilitarian fairness criteria in addition to RBRs, but not employer decision rights. When their utilitarian and RBR objectives conflict – as when

they experience a change in the experiment's *race* treatment – we estimate that members of Group 2 place about equal weight on these two objectives.

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.[48] 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 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. Our results suggest that this group of respondents is likely to resist policies that interfere with employers' decision rights, 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 conservative Americans might be more receptive to policies that equalize racial opportunities. One such situation is where a clear rule has *not* been applied in a race-blind way; in these cases, *restoring* race-blindness should have broad appeal given our results. Second, we show that respondents of all political leanings react more negatively to race-based actions that were taken for less-justifiable reasons, like personal animus and low-quality evidence. Antidiscrimination policies that target these types of behaviors may thus be better received than other policies.

[48] In this sense 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.

Our results in this paper are subject to some important *caveats* and leave some important questions unanswered. One important *caveat* is that all our results are *limited to the range of actions our scenarios depict*. Thus, for example, it seems likely that more *consequential* discriminatory actions (like being fired from a job or convicted of a crime), and less *justifiable* actions (such as ones based on racial hatred) would probably elicit stronger negative responses from respondents than we see. We might also see stronger, negative reactions, for example, to hiring scenarios in which the discriminatee is *more* qualified than his co-applicant. (We restrict attention to equally qualified applicants). Another limitation is that our scenarios are confined to a particular type of firm: sole proprietorships. We chose this context because it ensures that the recruiter has total control of the hiring decision and experiences its full financial consequences.[49] In larger firms, recruiters might not bear the full costs of indulging their own tastes or using lower-quality information. The strong and widespread support we see for *employer rights* among our respondents might also be more muted when the recruiter is an employee of a large firm.

While we have more than enough statistical power to test our pre-registered hypotheses, we also acknowledge that we lack statistical power to answer two important questions. First, does the discriminatee race effect really reverse sign (relative to the sample as a whole) among White conservatives? If White conservatives truly object more strongly to anti-White and anti-Black discrimination, this would complicate our description of conservatives, in general, as valuing race-blindness. Second, given our small sample, our data cannot shed much light on which factors explain non-White respondents' fairness assessments. Finally, we remind readers that the fairness assessments we elicit are not necessarily the same as the *actions* our respondents might take in real-world situations similar to our scenarios. For example, a business owner might choose to accommodate the discriminatory tastes of her customers while still experiencing that action as unfair.[50] That said, given the extensive evidence that people value fairness (e.g. Card et al. 2012; Cullen and Perez-Truglia 2018; Dube et al,

[49] In this respect, we follow Becker's (1971) classic exposition of employer taste-based discrimination: Becker's 'employers' make all of a firm's decisions (including hiring) and receive all the profits generated from the firm's operations. Assigning fairness ratings to our scenarios would be both more complex and more interesting if, for example, recruiters are balancing their personal assessments of what is best against company policies.

[50] That said, we note that on average our respondents rated this scenario as slightly more fair than unfair, suggesting that our respondents' real-world actions might indeed coincide with their fairness assessments in this case.

2019) our paper quantifies, for the first time, the fairness *costs* our subjects associate with taking different types of discriminatory actions.

Given all the above limitations, we view our results in this paper as a first step in understanding when and why ordinary people view discriminatory actions as unfair. One of many questions that could fruitfully be addressed in extensions of our work is the effects of the discriminatee's *individual* income (and opportunities) on respondents' fairness assessments. (Respondents' reactions to rich discriminatees from low-income groups could shed further light on utilitarianism, for example.) Other applications include different contexts, such as housing markets, credit markets, and judicial decisions; different discriminatee groups (such as gender, age, sexual orientation, political orientation, age, criminal and credit history); different social and psychological contexts *in the scenario* (for example, is the hypothetical action seen by hypothetical observers?; is the act depicted as conscious versus unintended?); different decision environments *for the respondent* (such as priming, cognitive depletion, audience effects, and personal exposure to previous discriminatory actions); and discrimination that is *embedded in laws and institutions* (as opposed to an individual's actions).

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.

Becker, Gary S. 1971. *The Economics of Discrimination* (second edition) Chicago: University of Chicago Press.

Bertrand, Marianne, and Esther Duflo. 2017. Field Experiments on Discrimination. In Banerjee, Abhijit Vinayak and Esther Duflo, eds. *Handbook of Economic Field Experiments*, vol. 1. Chapter 8 (pages 309-393) Also available as: NBER Working Paper 22014, 2016.

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.

Cappelen, Alexander W., Ranveig Falch and Bertil Tungodden 2019 "The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind" NHH Department of Economics Discussion Paper No. 06/2019.

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.

Everett, Jim A. C., Nadira S. Faber, and Molly Crockett. (2015). Preferences and beliefs in ingroup favoritism *Frontiers in Behavioral Neuroscience*, volume 9.

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.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. (2021) Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54. pages 1643–1662.

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

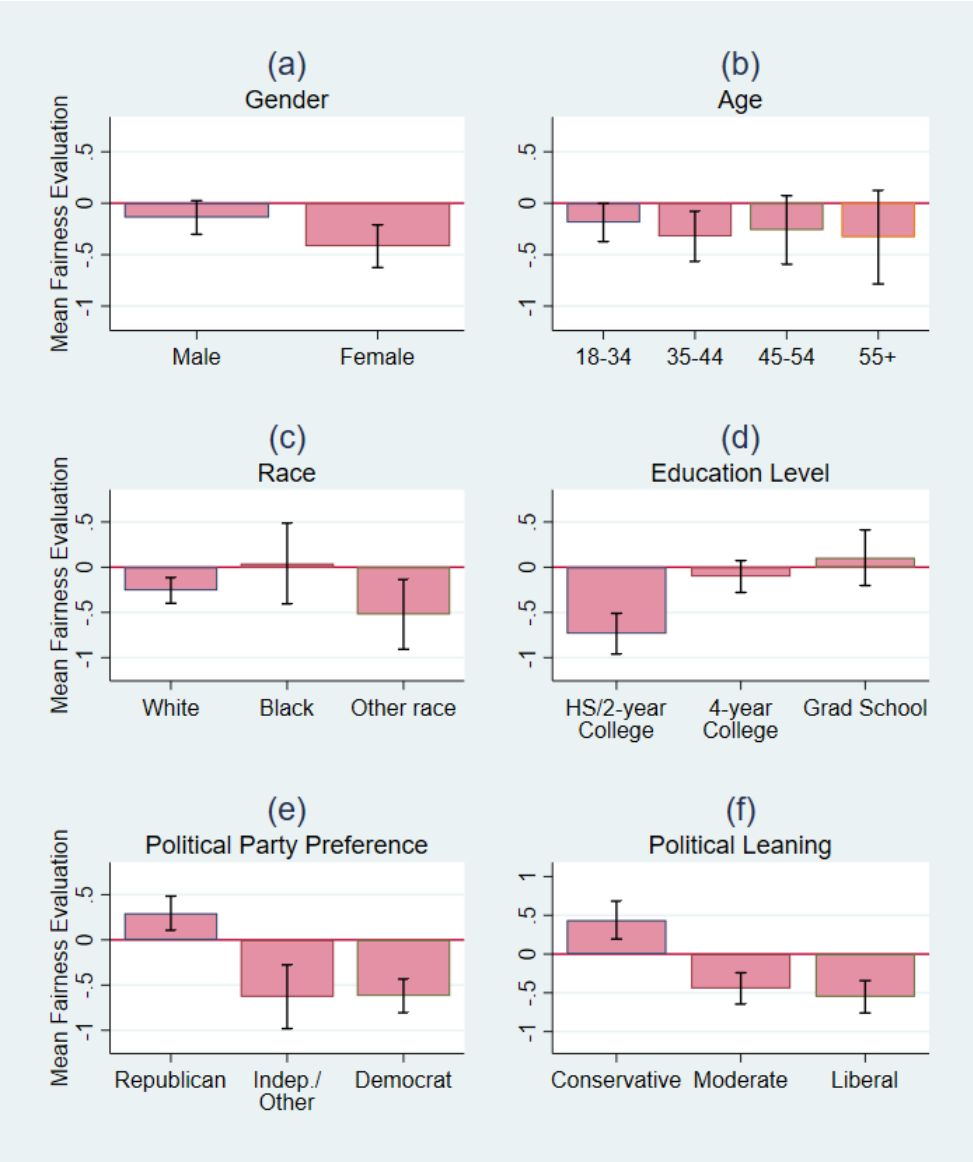
Schildberg-Hörisch, Hannah, Marco A. Schwarz, Chi Trieu, and Jana Willrodt 2022 "Perceived Fairness and Consequences of Affirmative Action Policies" CESifo working paper no.10198

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.

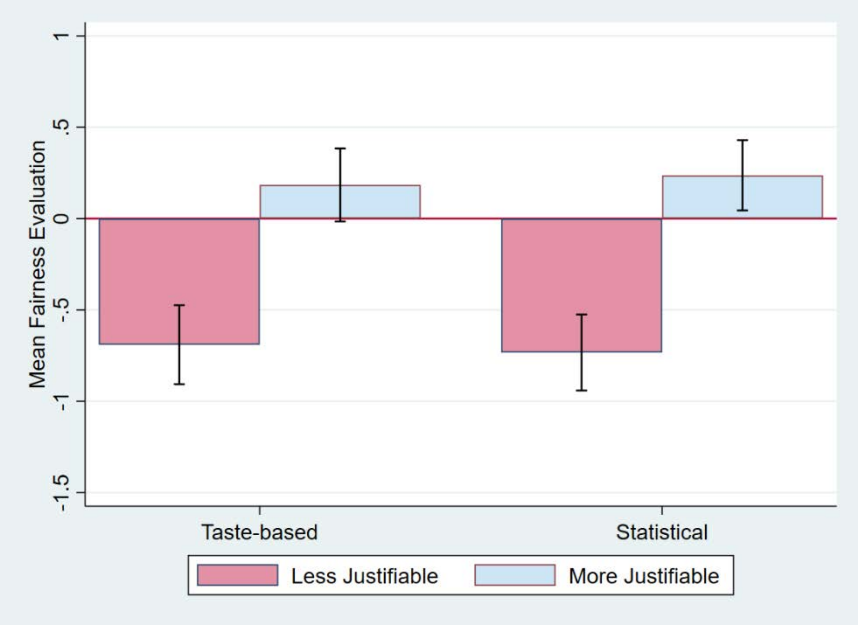
Figures

Figure 1: Mean Fairness of Discriminatory Actions by Respondent Characteristics



Notes: Fairness is measured on a scale from -3 ("very unfair") to 3 ("very fair"), where 0 was "neither fair nor unfair." This figure is based on only Stage 1 observations. 95% confidence intervals are shown. The *p*-values below are clustered by respondent.		
a) Gender: Males vs. Females = 0.037 **b) Age:** Ages 18-34 vs. 35-44 = 0.368 Ages 35-44 vs. 45-54 = 0.766 Ages 45-54 vs. 55+ = 0.805 Ages 18-34 vs. 55+ = 0.560	**c) Respondent Race:** White vs. Black = 0.204 Black vs. Other = 0.058 White vs. Other = 0.197 **d) Education level:** HS/2-year College. vs. 4-year College = 0.000 4-year College vs. Grad School = 0.246 Grad school vs. HS/2-year College = 0.000	**e} Political party preference** Republicans vs. Independents = 0.000 Independents vs. Democrats = 0.961 Democrats vs. Republicans – 0.000 **f) Political leaning:** Conservatives vs. Moderates = 0.000 Moderates vs. Liberals = 0.463 Liberals vs. Conservatives = 0.000

Figure 2: 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: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. *p*-values are clustered by respondent.

Figure 3: 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

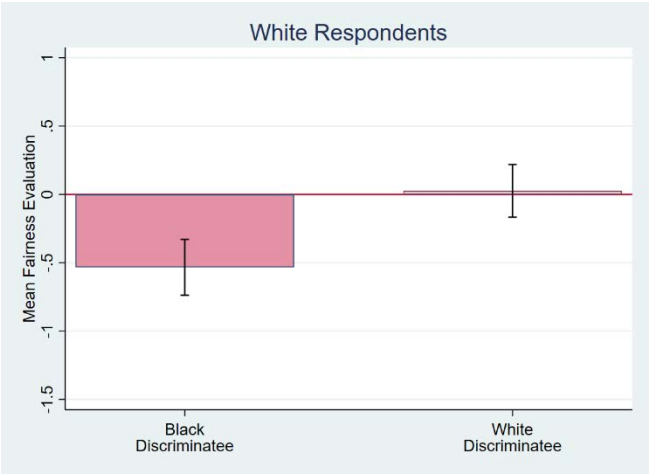
Within White Discriminatees: p=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. *p*-values are clustered by respondent.

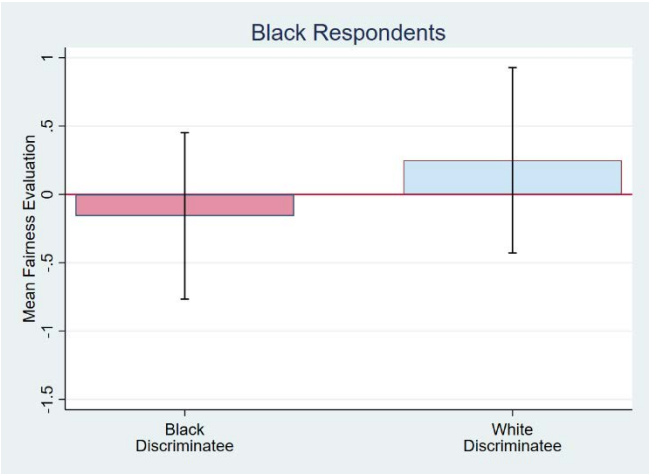
Within Black Discriminatees, less-justifiable scenarios are 0.898 units less fair. Within White Discriminatees, less-justifiable scenarios are 0.953 units less fair. A test for equality of the Less versus More *Justifiability* Gap between the Black and White treatment yields p = .679.

Figure 4: Fairness Ratings by Respondent Race and Discriminatee Race

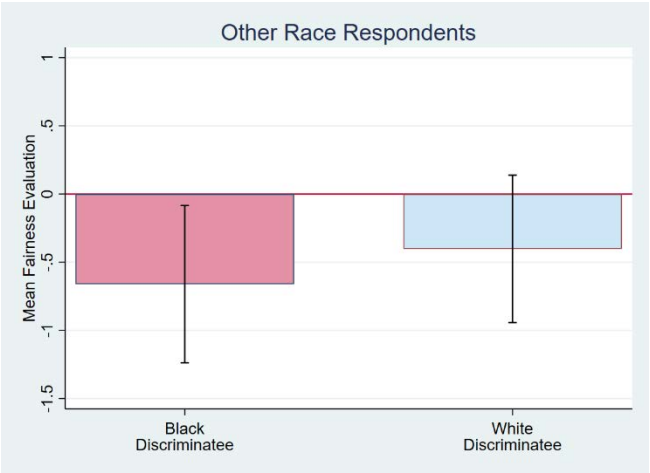
(a)



(b)



(c)



p-values:

Black versus White Treatment:

Overall: Overall: p=.000

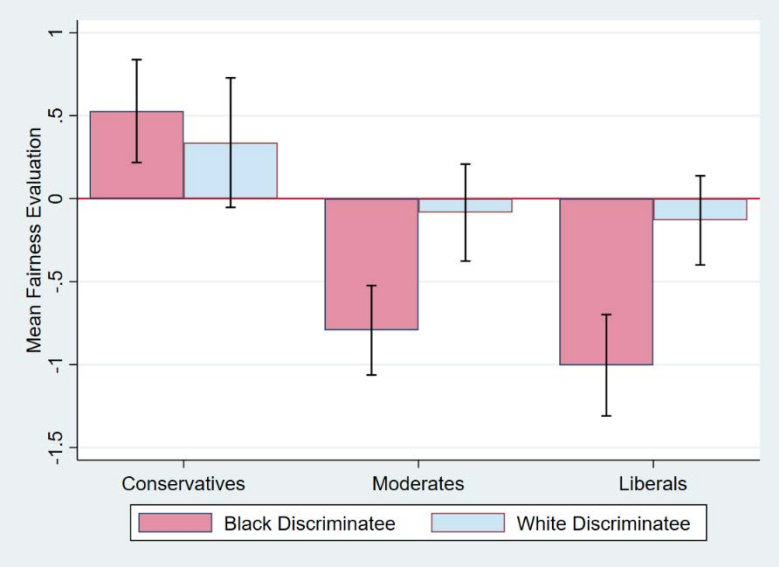
Within White respondents: p=.000

Within Black respondents: p=.365

Within Other respondents: p=.506

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 5: Fairness Ratings by Political Orientation and Discriminatee Race



p-values:

Black versus White Treatment:

Overall: *p*=.000

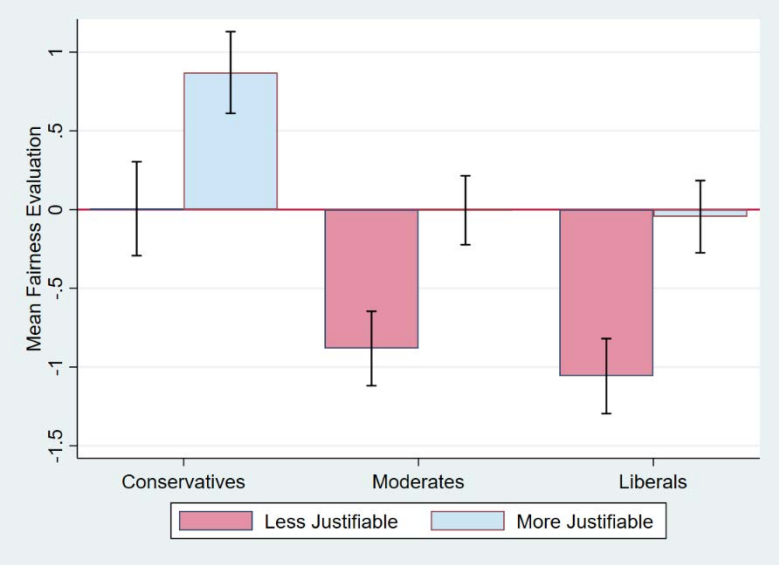
Within Conservatives: *p*=.448

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 6: 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

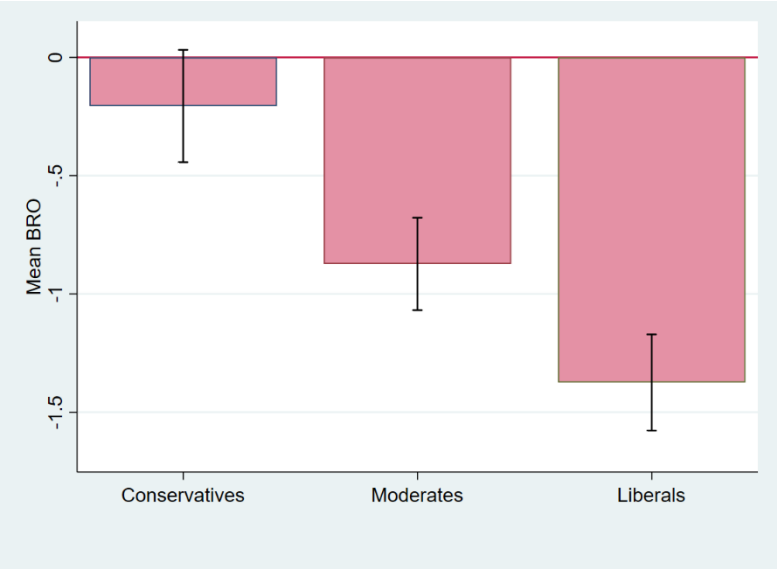
Within Conservatives: *p*=.000

Within Moderates: *p*=.000

Within Liberals: *p*=.000

Note: Figure is based on Stage 1 observations only. 95% confidence intervals are shown. 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 7: Respondents' Perception of Black Peoples' Relative Economic Opportunities (BRO) by Political Leaning



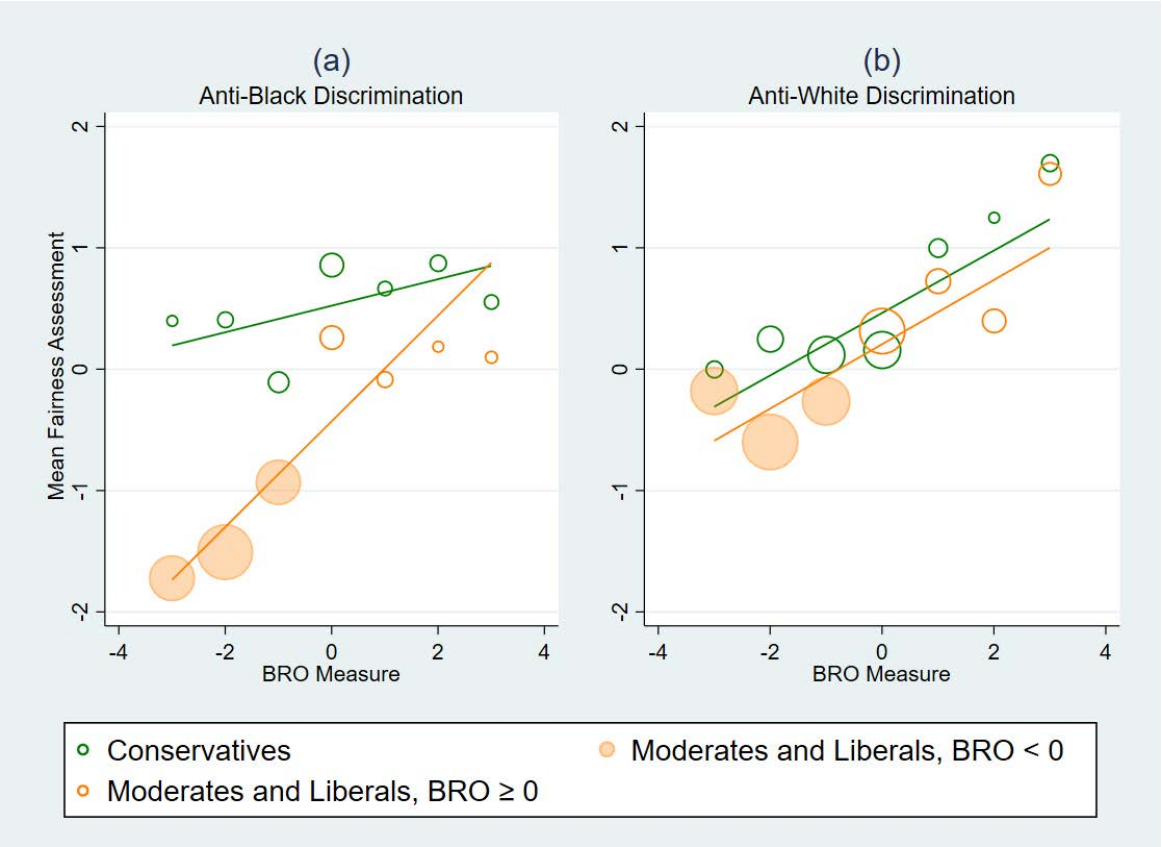
p-values:

Conservatives vs. Moderates: *p*=.000
Moderates vs. Liberals: *p*=.001
Liberals vs. Conservatives *p*=.000

Note:

BRO is the respondent's assessment of Black peoples' relative economic opportunity on a scale of -3 (much less) to 3 (much more). Figure is based on Stage 1 observations only. 95% confidence intervals are shown. All *p*-values are clustered by respondent. A test for equality of BRO across all three political groups yields $p = .577$.

Figure 8: Political Differences in Fairness Ratings, by Perceived Relative Opportunities (BRO) and Discriminatee Race



Note: 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.

- 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 ($p = .000$)