Extreme justifications fuel polarization*

Christiane Buschinger[1] Markus Eyting[1,2] Florian Hett[1]
Judd Kessler[3]

First version: February, 2025
Last updated in May, 2025

[1] Johannes Gutenberg University Mainz
[2] Leibniz Institute for Financial Research SAFE
[3] The Wharton School of the University of Pennsylvania

Abstract

How does polarization — as measured by mistreatment of political rivals — spread? In an online experiment, participants choose between splitting financial resources equally or discriminating against a member of the opposing political party. We vary the information subjects receive about others' choices and justifications for discrimination. Exposure to extreme justifications for discrimination increases discrimination — particularly in a polarized environment, when many others are already discriminating — and it leads participants to adopt more extreme justifications themselves. Our findings suggest a self-reinforcing dynamic that may fuel polarization: Exposure to extreme statements increases polarization and the prevalence of extreme reasoning.

Keywords: political polarization, peer effects, justifications, outgroup discrimination, social norms

JEL codes: C9, D01, D9

*Address for correspondence: Judd B. Kessler; 320 Dinan Hall; 3733 Spruce Street; Philadelphia, PA 19104 USA. We thank the Leibniz Institute for Financial Research SAFE, the Wharton School, and the Wharton Behavioral Lab for funding. The project received IRB approval at the joint ethics board of Goethe University Frankfurt and JGU Mainz.

1 Introduction

An active and growing literature in the social sciences documents the prevalence and intensification of discrimination along partisan lines, fueling what is commonly referred to as polarization (Iyengar and Westwood, 2015; Boxell et al., 2024; Lane et al., 2024). In the United States, Democrats and Republicans increasingly express dislike, display distrust, and even engage in mistreatment toward each other (Chen and Rohla, 2018; Dimant, 2024). It is thus important to understand the process that drives these dynamics of polarization.

In this paper, we explore the decision to discriminate against a political outgroup and identify a force that might contribute to an increase in polarization over time: the use of extreme justifications. We test whether extreme justifications repel or attract observers toward discriminatory actions and how they shape individuals' own justifications of their actions.

Our experimental design builds upon the well-established bystander allocation game (Tajfel et al., 1971; Chen and Li, 2009; Kranton et al., 2020). In our study, a subject — either a Democrat or Republican — allocates money between two other participants: one Democrat and one Republican. Subjects can choose to split the amount equally or to allocate the entire sum to the member of their own political party.

Treatments vary whether or not subjects observe the decisions of other participants, who we call "peers." Peers are all members of the subject's own political party. In treatments in which peer behavior is shown, we randomly vary how many peers engage in discrimination against the political outgroup. We also vary whether or not subjects see a justification for discrimination provided by one (and only one) of the peers. In the *No Justification* treatment, subjects see peer behavior but are not provided with a justification for discrimination. In the *Moderate Justification* treatment, they are shown a socially acceptable, moderate justification for discrimination (e.g., one that indicates being more aligned with the political ingroup). In the *Extreme Justification* treatment, they are shown a socially inappropriate, extreme justification for discrimination (i.e., one that calls the political outgroup a "cancer to this country").

Relative to being shown no justification or a moderate one, being exposed to an extreme justification encourages discrimination. This effect of extreme justifications is driven by environments where many other peers are discriminating, even though the justification is provided by only one peer.

Finally, we investigate how individuals' own justifications for discrimination are influenced by the justifications they observe, and we find that exposure to extreme justifications increases the likelihood that individuals adopt similar reasoning themselves.

Interestingly, however, the effect on justifications is subtle. Exposure to extreme justifications does not change participants' perceptions of the social appropriateness of the

statements: participants still evaluate these extreme justifications as socially inappropriate. It also does not drive an increase in the fraction of subjects who use an extreme justification as their primary reason for discriminating. Instead, exposure to an extreme justification leads subjects to be more likely to list an extreme statement among a list of justifications that align with their views. They understand that the justification is socially inappropriate, but they become willing to adopt it after observing it being used by someone else.

By exploring how exposure to extreme justifications can amplify partisan discrimination, our paper contributes a potentially new mechanism to the literature on political polarization (Iyengar and Westwood, 2015; Dimant, 2024; Lane et al., 2024), social identity (Tajfel et al., 1971; Chen and Li, 2009; Charness and Chen, 2020; Shayo, 2020), and their interaction (Iyengar et al., 2012; Iyengar et al., 2019). Our work also relates to the recent literature on discrimination, which increasingly explores mechanisms beyond the classical distinction of statistical and taste-based discrimination (Bohren et al., 2019; Bohren et al., 2022; Bohren et al., 2025; Eyting, 2022). While most work in these areas considers various forms of behavior and beliefs and their underlying social dynamics, we consider the role of statements (Kessler, 2017). This connects to recent work on the role of language and framing in polarization, like Djourelova (2023) and Bursztyn et al. (2023) which show that restricting extreme or slanted language can reduce behavioral polarization.

Our work also contributes to the extensive literature on peer effects. A key finding across this literature is that individuals' behavior is strongly influenced by what they observe others doing. Keizer et al. (2008) provides field evidence for the "broken window" theory (Wilson and Kelling, 1982), showing that minor norm violations can trigger broader norm erosion. These findings underline that the mere existence of norm violations — even if not yet widespread — can suffice to shift behavior, a mechanism our results corroborate: observing only one extreme justification among peers increases discrimination. Complementary, studies like Carrell et al. (2008) demonstrates that unethical behavior, such as academic cheating, spreads among peers. Dimant (2019) highlights an important asymmetry in such peer effects: anti-social behaviors are more contagious than pro-social behaviors, and social proximity amplifies contagion. Relatedly, Bursztyn et al. (2020) finds that when xenophobic actions are perceived as common and tolerated, individuals are less likely to sanction them. Our findings confirm the general existence of peer effects in anti-social behavior, while also documenting a nuanced interaction with their justification. In addition, while subjects only play our game once, our finding that being exposed to an extreme justification leads others to adopt more extreme justifications suggests the possibility of these polarized views spreading over time (Bisin and Verdier, 2001; Bisin and Verdier, 2011).

The literature on social norms considers a conceptual distinction between descriptive

norms (what most people do) and injunctive norms (what is considered appropriate). Bicchieri and Xiao (2009) shows that descriptive norms, rather than injunctive norms, are the primary drivers of behavior in social dilemmas. Bicchieri et al. (2022) and Gächter et al. (2017) further show how peer behavior influences norm perceptions and compliance. Barr et al. (2018) shows that discrimination is moderated by the perceived social inappropriateness of discriminatory actions. Our study extends these findings by showing that exposure to extreme justifications operates as a somewhat distinct mechanism: Justifications matter conditional on the behavior of others (i.e., the descriptive norm). In addition, we do not observe a change in the injunctive norm in the form of an adjustment of what people think is socially appropriate: subjects continue to view extreme justifications as socially inappropriate but still discriminate more after exposure.

2 Experimental Design and Data

Our data was collected in two waves (in December 2024 and March 2025) from an online artefactual field experiment, programmed in Otree (Chen et al., 2016) and run on Prolific. Each wave of the experiment was pre-registered in an OSF repository (Eyting et al., 2024).

2.1 Experimental Design

The experiment proceeded in several stages.[1] First, subjects created a personalized avatar. Second, they learned about other members of their own party, their "peer group," who — in a previous session — faced the same decision they were about to make. Third, subjects received information about their peers' choices, the specific details of this information depending on treatment. Fourth, subjects made a decision about whether or not to discriminate against a political outgroup; along with their decision of whether to discriminate, subjects were asked to provide a justification for their choice from a set of six pre-determined statements (the options differed based on whether the subject opted to discriminate). Fifth, we elicited subject beliefs about their peers' behavior. Sixth, we asked subjects to select any additional reasons for their choice with which they agreed. Finally (in wave 2), we elicited perceptions of social norms by asking subjects about the social appropriateness of the six justification statements. The study concluded with a short questionnaire that elicited demographic data and asked an attention-check question.

Below, we provide additional details about the stages of the experiment that do not vary by treatment. We then describe the experimental treatments and the stages of the experiment that vary by treatment.

[1] Screenshots of all instructions and pages in the experiment can be found in Appendix A.1.

Avatar Creation At the beginning of the experiment, all subjects were invited to create an avatar that resembled them (choosing from various skin tones, hairstyles, hair colors, clothing, and glasses). The idea behind the avatar creation was to help subjects who saw avatars of other participants think about them as representing real people.

Peer Group All subjects were then introduced to their "peer group," which consisted of 10 participants from a previous session who share the political preferences of the subject (i.e., were members of the same political party).[2] The peers were shown by displaying their self-created avatars.

The information that subjects received about the members of their peer group varied by treatment as described in Section 2.2.

Allocation Decision and Justifications All subjects were then asked how they wanted to allocate $10 in a binary other-other allocation game. Subjects could either split the money equally between a randomly chosen Democrat and a randomly chosen Republican or allocate the full amount to the subject who was a member of their political party.

This choice is the main outcome of interest in the study. Choosing to discriminate against a member of the opposing political party in favor of one's own political party in a setting where an equal split of resources is feasible (and is typical in allocation games where subjects are anonymous) is our measure of polarized behavior.

After making this decision, all subjects are also asked to provide a reason for their choice. They could choose one option from a list of six pre-determined statements that — in the case of the decision to discriminate against the political out-group — ranged from moderate to extreme.[3] The choice of justification is the second main outcome of interest in our study. At this stage, subjects were asked for only one justification. Later in the study, we invited subjects to select all the justifications with which they agreed.

Perceived Social Appropriateness After subjects selected all the justification with which they agreed, subjects from wave 2 participated in a norm elicitation task in the spirit of Krupka and Weber (2013).[4] In particular, we told subjects that 100 prior participants who were members of their political party were asked to rate each of the six justification statements as either socially appropriate or socially inappropriate. We asked subjects to guess how many of those 100 said the statement was socially appropriate

[2]Subjects' political preferences are obtained from Prolific screener information and validated using a survey question at the end of the study.

[3]See Appendix Table B1 for an overview of the statements subjects could select. We externally validated the extremity of each statement in a separate Prolific study. See Appendix A.3 for details and results of this *Validation Study*.

[4]We added this in wave 2 after the pattern of results in wave 1 caused us to wonder whether subjects thought that extreme statements were more socially appropriate after observing them.

and gave them a financial incentive for guessing a range that included the correct answer.[5] This question allows us to measure subjects' perceptions of whether the justifications are socially appropriate.

2.2 Treatments

In each of the waves, subjects were randomized to one of four treatments, which are described below. The treatments differ in what information subjects received about their peers' decisions of whether to discriminate and their justification for doing so.

Baseline Treatment In the Baseline treatment, participants saw the 10 avatars of their peer group members but received no information about their peers' allocation decisions or any information on their justifications. While not used in the main analysis, the Baseline treatment provides a benchmark level of discrimination and provides data on prior beliefs about the number of peers who discriminate, which is useful information for deciding how to proceed with our analysis, as discussed below.

Peer Information Treatments Subjects who were not randomized into the Baseline treatment were randomized into one of three peer information treatments. In these treatments, subjects were provided with information about the allocation decisions of their peers.

In particular, subjects saw n peers who had discriminated against the political outgroup and $10-n$ who split the money equally between the ingroup and outgroup members. Different subjects saw a different number of discriminators, n, with n ranging from 1 to 10. In all cases, we showed a first peer who chose to discriminate, and introduced variation in the choices of the other 9 peers (who may have discriminated or not discriminated).[6] Subjects saw each peer's allocation decision displayed as a sentence displayed above the respective avatar in a "speech bubble." As shown in Appendix Figures A9–A11, Subjects saw speech bubbles that said either: "I gave all $10 to the Democrat" (Republicans saw the world "Republican" instead of "Democrat") or "I split the money equally between the two".

The provision of peer information was the same across all three peer information treatments. What differed across them was whether a justification was provided by the

[5]The 100 participants that comprised this reference group came from the *Validation Study* as described in Appendix A.3.

[6]In wave 1, we randomly selected peers from prior participants who had taken the study. Since 52% of participants who were eligible to be peers had discriminated, this random selection process meant many subjects saw a nearly even number of peers who discriminated and who did not, and more-extreme distributions were rare. In wave 2, we over-sampled from these extremes to create a more uniform distribution of peers who discriminated, which allowed us to explore behavior among subjects who learned that $1-3$ or $8-10$ subjects discriminated, cases that came up rarely in wave 1.

first avatar in the set of ten peers and whether that justification was a moderate, socially acceptable statement or an extreme, socially inappropriate statement.[7]

No Justification Treatment In the No Justification treatment, subjects saw the peer information but no justification for discrimination from the first peer (see Appendix Figure A9).

Moderate Justification Treatment In the Moderate Justification treatment, subjects saw a justification for discrimination from the first peer that was presented in the speech bubble after their choice, in bold letters. For Democrats it said: "I agree more with the values and morals of the Democrats". Republicans saw the same statement with the word "Democrats" replaced by "Republicans" (see Appendix Figure A10). This statement was deemed the most moderate and the most socially acceptable of the six statements by participants in an accompanying validation study taken by different people from the same subject pool.[8]

Extreme Justification Treatment In the Extreme Justification treatment, subjects saw a different justification for discrimination from the first peer that was presented in the speech bubble after their choice in bold letters. For Democrats it said: "Republicans are a cancer to this country and should not be supported in any way." Republicans saw the same statement with the word "Republicans" replaced by "Democrats" (see Appendix Figure A11). This statement was deemed the most extreme and the least socially acceptable of the six statements by participants in the accompanying validation study taken by different people from the same subject pool.[9]

Beliefs About Others' Allocation Decisions After subjects made their choices, we elicited their beliefs about how many people in their peer group distributed money equally

[7]In wave 1, we randomly selected a first avatar among participants in a pre-study run before wave 1 who chose the moderate or extreme message as their primary justification. In wave 2, we chose subjects in wave 1 who had agreed to both extreme and moderate justifications from the full list. Subjects were then shown one of these avatars as their first peer, paired with either no justification, a moderate justification, or an extreme justification. This approach varied only the message, not the avatar image. In wave 1, we also randomized some subjects to have slightly different instructions. We did this by replacing the information about peers that reads: "You can see their decisions below, but they will never see your decision" on the treatment page with the text: "You can see their decisions below. In one of the following sessions your avatar and decision will be at the top left, and thereby prominently shown to another participant." This version was introduced in an attempt to induce social image concerns. However, the results did not differ systematically, presumably because the treatment was too weak. As a result, we pool the data from both versions of the instructions in our analysis and include a dummy in our regressions for whether you were shown the alternative sentence interacted with the number of peers who discriminate. We dropped this variation in wave 2.

[8]Appendix Figures B1 and B2 show the mean ratings for all possible justifications. This moderate justification is called "Values" in the figures.

[9]This extreme justification is called "Cancer" in Appendix Figures B1 and B2.

between the Democrat and Republican and how many chose to provide all $10 to their political ingroup member.

In the Baseline treatment, these predictions represent prior beliefs regarding monetary allocations of others, as no peer information is actually shown. For this treatment we offered a $0.50 bonus payment if the subject guessed the correct answer. As shown in Appendix Figure B3 we see a mean prior belief that 5.75 peers discriminated, which is close to the truth.

In the three other treatments, participants actually saw the number of peers who discriminated, which makes this elicitation more of a manipulation check. Although we did not provide an incentive for a correct answer, 26% of the subjects reported the right number, 57% were within 1 of the truth, and 74% were within 2 of the truth (rates do not systematically vary by treatment). The correlation coefficient between the peers shown discriminating and the number reported by subjects is high (0.46, $p < 0.001$). So, while it seems most subjects did not count the number of peers who discriminate and report it back to us, subjects clearly perceived more discrimination in environments when they were shown more peers who indeed discriminated.

3 Results

A total of 2587 participants completed the experiment.[10] The study took about 5 minutes to complete in wave 1 and 8 minutes in wave 2. Section A.2 reports the minor adjustments to the design we implemented in wave 2. Subjects earned on average $0.90 in wave 1 and $1.40 in wave 2, which both correspond to an hourly wage of around $10.50. The payment consists of the participation fee and a potential bonus payment of up to $1.00. We exclude 96 subjects for whom the Prolific screening information on political orientation did not align with the information given by the subject in our survey. We also exclude the 23 subjects who failed the attention check in the survey.[11] Neither of these exclusions vary by treatment. These exclusions leave us with a final sample of 2471 subjects. These subjects are equally split by political party (Democrat and Republican) and by gender.[12]

3.1 Effect of Justifications on Discrimination

Is discrimination impacted by exposure to different justifications? To assess the impact of justifications, we compare choices across the three peer information treatments. Figure 1 compares levels of discrimination across the three treatments (the level of discrimination in the Baseline treatment is shown by the dashed line).

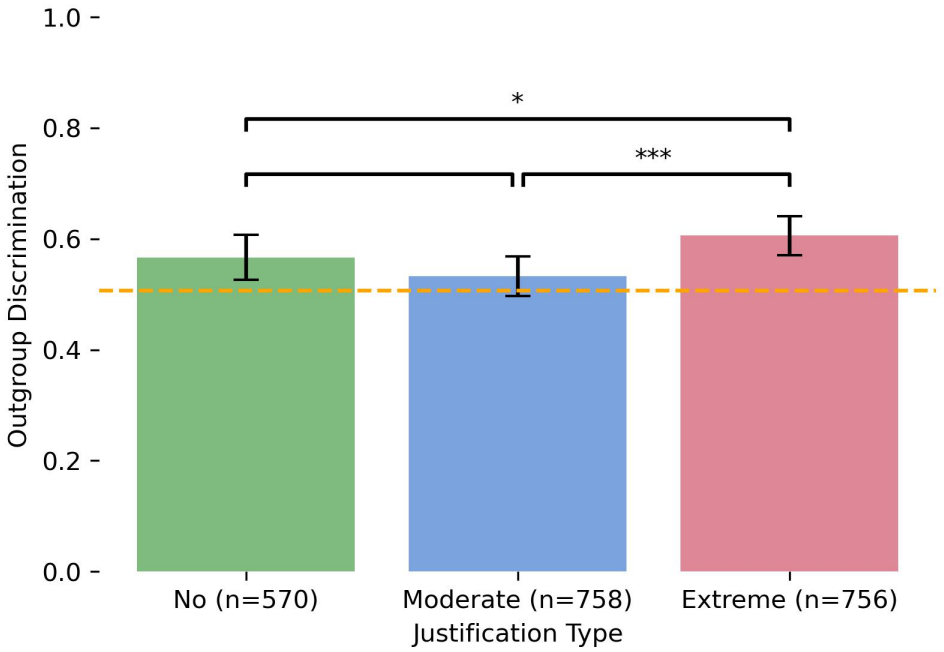
[10] We pre-screened participants based on their location (US) and first language (English) and excluded participants who had participated in the pre-study.

[11] We excluded participants who failed to select the second option from a list when instructed to do so.

[12] See Appendix Tables B2 and B3 for descriptive statistics and balance checks.

Comparing the levels, it looks as though exposure to an extreme justification generates a higher discrimination rate than observing no justification or than being exposed to a moderate justification. On average, 60.6% of subjects discriminate when exposed to an extreme justification. This number is 56.67% for subjects who do not observe a justification and only 53.30% for those who are exposed to a moderate justification.

Figure 1: The Effect of Justifications on Discriminatory Behavior



Notes: "Outgroup Discrimination" is a binary variable equal to 1 if the unequal split was chosen. The green, blue, and red bars show the likelihood of discrimination across the "No Justification," "Moderate Justification", and "Extreme Justification" treatments, respectively. The dashed yellow line shows the level of discrimination in the Baseline treatment. Significance stars are based on the regression specification in column (1) of Table 1. The vertical error ranges represent 95%, confidence intervals. *p $<$ 0.1, **p $<$ 0.05, ***p $<$ 0.01.

The cleanest variation in our study comes from subjects who participated in the same wave, saw the same instructions, and were randomized to the same number of discriminating peers but who were randomly exposed to different justifications for discrimination. To leverage this variation only, our regression analysis always controls for dummies that interact dummies for the number of peers who discriminate with dummies for wave and the instructions received in wave 1 (as described in footnote 7).

Column (1) of Table 1 presents the results of statistical tests using this regression specification. It finds that the differences in discrimination between the Extreme Justification treatment and the No Justification and Moderate Justification treatments are statistically significant at ($p = 0.067$ and $p = 0.001$, respectively), as also shown with significance stars in Figure 1.[13]

[13]This main effect is not isolated to one political party. Appendix Table B4 shows that this pattern arises among both Democrats and Republicans, with the initial level of discrimination slightly higher for Democrats and the treatment effect of the the extreme justification on discrimination directionally larger for Republicans.

Table 1: The Effects of Justifications on Outgroup Discrimination

Variables	Pooled	≤ 5 Peers	≥ 6 Peers
	(1)	(2)	(3)
Extreme Justification	0.051*	0.014	0.086**
	(0.028)	(0.041)	(0.039)
Moderate Justification	-0.032	-0.040	-0.024
	(0.028)	(0.041)	(0.039)
No Justification Mean	0.567	0.543	0.588
Extreme Justification – Moderate Justification			
Difference	0.083	0.054	0.110
p-value	0.001***	0.155	0.001***
Wave × No. of Peers Dummies	X	X	X
Instructions × No. of Peers Dummies	X	X	X
Observations	2084	987	1097

Notes: This table presents regression results on discrimination controlling for dummies for wave and instructions in wave 1, each interacted with dummies for the number of peers who discriminated. The "No Justification" treatment is the excluded group. Column (1) shows results for the main sample in the "No Justification," "Moderate Justification," and "Extreme Justification" treatments. Column (2) shows results for the subset of these subjects who were randomly chosen to be shown ≤ 5 peers discriminating and column (3) shows results for subjects who were randomly chosen to be shown ≥ 6 peers discriminating. *p< 0.1, **p< 0.05, ***p< 0.01.

Result 1 – The Effect of Justifications: *Seeing an extreme justification increases discriminatory behavior relative to seeing no justification and to seeing a moderate justification.*

3.2 The Role of Peers' Choices on the Effect of Justifications

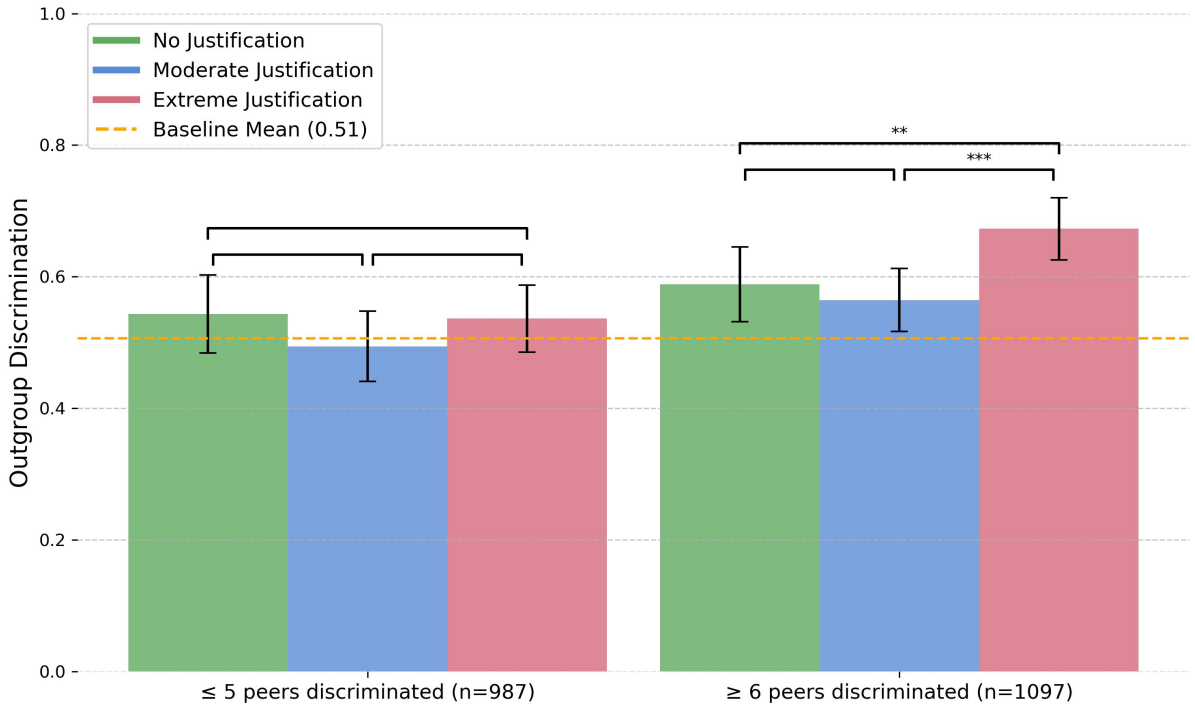
Next, we take a closer look at how the effectiveness of justifications differ by peer behavior. In Figure 2, we split subjects based on whether they observed between 1 and 5 peers discriminate or between 6 and 10 peers discriminate.[14]

Splitting the results this way proves fruitful. Figure 2 shows that the effect of justifications on discrimination is driven by environments in which subjects observe the majority of peers discriminating. In these settings, discrimination is higher in the Extreme Justification treatment than in the No Justification treatment (67.27% vs 58.84%) and higher than in the Moderate Justification treatment (67.27% vs 56.46%). Meanwhile, when

[14]There are several reasons to choose this split to explore heterogeneity in treatment effect by peer behavior. First, it divides the support of the number of peers who discriminate equally into two blocks of five. Second, it constitutes the median split in the number of peers that subjects saw discriminating. Third, the average prior belief of how many peers discriminated is between 5 and 6, as shown in Appendix Figure B3. Nevertheless, results look very similar regardless of where we split the data in terms of the number of peers discriminating (as shown in Appendix Figures B4 and B5). For more information on the discrimination rate across the number of peers who discriminate in different treatments, see Appendix Figure B6.

subjects observe five or fewer peers discriminating, there is no impact of an extreme justification on discrimination. In contrast to extreme justifications, moderate justifications do not increase discrimination over no justification. If anything, moderate justifications directionally reduce discrimination relative to providing no justification.

Figure 2: The Effects of Justifications on Discriminatory Behavior across Others' Behavior



Notes: "Outgroup Discrimination" is a binary variable equal to 1 if the unequal split was chosen. The green, blue, and red bars show the likelihood of discrimination across the "No Justification," "Moderate Justification," and "Extreme Justification" treatments, respectively. The dashed yellow line shows the level of discrimination in the Baseline treatment. The three bars on the left include the subset of subjects who were randomly chosen to be shown ≤ 5 peers discriminating and the three bars on the right show results for subjects who were randomly chosen to be shown ≥ 6 peers discriminating. Significance stars are based on the regression specifications in columns (2) and (3) of Table 1. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01.

Columns (2) and (3) of Table 1 show regression results using the same specification in column (1), as described above, but split the data based on the number of peers who discriminate. The results confirm that being exposed to an extreme justification for discrimination increases discrimination when many peers also discriminate. The effect of an extreme justification on discrimination is estimated to be only 1.4 percentage points in column (2) when 5 or fewer peers discriminate, but it is a statistically significant 8.6 percentage points ($p = 0.025$) when 6 or more peers discriminate.

The effect of observing a moderate justification is negative throughout, although estimates are not statistically significant. Consequently, relative to observing a moderate justification, observing an extreme justification increases discrimination by 8.3 percentage points overall ($p = 0.001$), and this effect is directionally larger (11 percentage points, $p = 0.001$) among participants who observed a majority of peers discriminating.

Result 2 – When do Extreme Justifications Work: *Extreme justifications seem particularly effective at increasing discrimination in environments when many peers are also discriminating.*

3.3 The Spread of Justifications

Does being exposed to an extreme justification for discrimination lead subjects to adopt the extreme justifications themselves? If so, this type of spillover would constitute an additional, indirect dynamic effect of extreme justifications on polarization.

To answer this question, we explore the justifications subjects give for their own discriminatory behavior and test whether they differ based on the justification to which they were exposed.

First, we explore the primary justification provided by subjects about why they choose to discriminate. We find that primary justifications are not affected by treatment. The left set of three bars in Figure 3 shows the probability of selecting the most extreme statement (i.e., the one shown to subjects in the Extreme Justification treatment) as a subject's *primary* justification. Rates of selecting that extreme justification are low and do not differ by treatment, even though more people discriminate in the Extreme Justification treatment than in the other two (for the same results looking only at subjects who discriminated, see Appendix Figure B7).

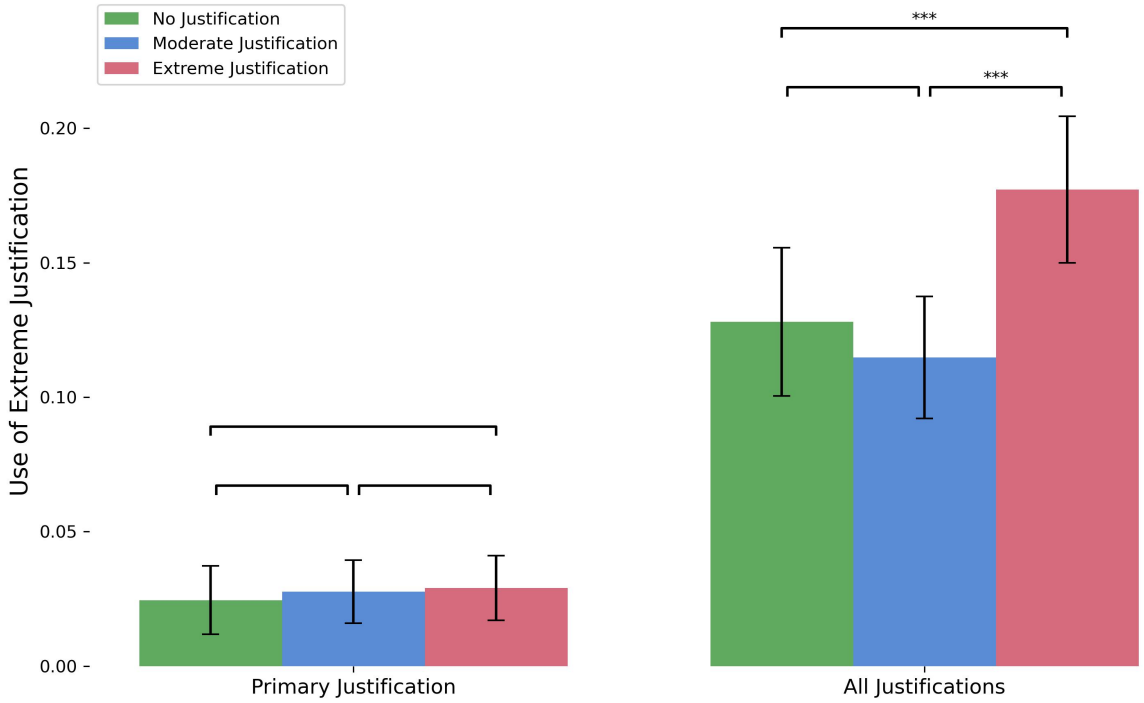
In addition, we do not see a significant difference in the use of a somewhat-less-extreme (i.e., "intermediate") statement or in the use of one of the two moderate statements as the primary justification (see Appendix Figure B8).[15]

While exposure to extreme justifications does not change the likelihood of selecting the extreme statement as a primary justification, it does increase the likelihood that subjects include it in their full list of reasons for discrimination. The right set of three bars in Figure 3 shows the likelihood that a subject selects the extreme statement as one of the statements with which they agree. We see that more subjects say they agree with that statement in the Extreme Justification treatment.

Table 2 confirms these results in a regression framework using the same regression specification as Table 1 but with the dependent variable of selecting the extreme justification. Column (1) confirms that there is no effect of treatment on the likelihood that the extreme statement is chosen as a primary justification. Column (2), on the other hand, shows that there is a 5.6 percentage point increase in agreeing with the statement ($p < 0.001$), which represents a 44% increase in the likelihood that the statement is selected on a base of 12.8 percentage points. Columns (3) and (4) condition on the subject discriminating and show the same pattern of behavior: no increase in use of the

[15]We condition on discriminating in this analysis since we have more people who discriminate in the Extreme Justification treatment and so we have more primary justifications overall in that treatment.

Figure 3: The Spread of Extreme Justifications



Notes: "Use of Extreme Justification" is a binary indicator equal to 1 if the most extreme statement was used as the primary justification (left three bars) or was selected as one of the justifications with which the subject agrees (right three bars). See Appendix Figure B7 for this same figure conditioning on individuals who discriminated. Significance stars are based on regression specifications as shown in columns (1) and (2) of Table 2. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01

extreme statement as a primary justification but an increase in indicating agreement with the statement. While the effect of the extreme justification on discrimination was particularly strong when the majority of peers discriminated, Appendix Figure B9 shows that the effect on the spread of extreme justifications looks very similar regardless of the number of peers who discriminated.

This increase in agreement with this extreme statement might come from simply adding the extreme statement to the list or from replacing another justification with the more extreme one. We see that subjects in the Extreme Justification treatment indeed provide directionally more justifications overall, 0.204 more per subject (conditional on discriminating). While this directional increase is only marginally statistically significant ($p = 0.094$), it could theoretically be responsible for the 7.6 percentage point increase in the fraction of subjects who report that they agree with the extreme statement, as estimated in column (4).

Finally, we investigate potential mechanisms for the increase in use of the extreme justifications. A plausible candidate is a shift of the perceived social appropriateness of extreme statements. Observing others justifying their behavior with an extreme statement might increase someone's own assessment on how acceptable the statement is to make, thereby encouraging its use. We find no such effect. Appendix Figure

Table 2: The Spread of Extreme Justifications

Variables	Unconditional		Conditional	
	Primary	All	Primary	All
	(1)	(2)	(3)	(4)
Extreme Justification	0.005	0.056***	0.006	0.076**
	(0.010)	(0.020)	(0.016)	(0.033)
Moderate Justification	0.004	-0.011	0.013	-0.000
	(0.009)	(0.019)	(0.016)	(0.032)
No Justification Mean	0.025	0.128	0.043	0.226
Extreme Justification – Moderate Justification				
Difference	0.001	0.067	-0.006	0.077
p-value	0.940	<0.001***	0.663	0.010**
Wave × No. of Peers Dummies	X	X	X	X
Instructions × No. of Peers Dummies	X	X	X	X
Observations	2084	2084	1185	1185

Notes: This table presents regression results on the use of the extreme justification controlling for dummies for wave and instructions you saw in wave 1, each interacted with dummies for the number of peers who discriminated. The "No Justification" treatment is the excluded group. The dependent variable in columns (1) and (3) is a binary variable for having chosen the extreme statement as the subject's primary justification. The dependent variable in columns (2) and (4) is selecting the extreme state as one of the justifications with which the subject agrees. Columns (1) and (2) show all data from the "No Justification," "Moderate Justification," and "Extreme Justification" treatments. Columns (3) and (4) show results conditioning on participants who discriminate. *p< 0.1, **p< 0.05, ***p< 0.01.

B10 shows that being exposed to an extreme justification does not increase the belief of how many others find extreme statements socially appropriate. This suggests that norm-shifting is unlikely to be the underlying mechanism.

Result 3 – The Spread of Justifications: *Observing an extreme justification increases the likelihood of adopting an extreme justification as a reason for discriminating. However, it does not increase the likelihood of using it as a primary justification for discriminating and it does not change beliefs about the social appropriateness of the statement.*

4 Conclusion

We investigate the impact of justifications in fueling political polarization as measured by the choice to discriminate against a political outgroup.

We document three key results. First, extreme justifications increase the likelihood of discriminatory behavior. Second, this holds especially in environments where such behavior is already prevalent. In contrast, moderate justifications, which are perceived as socially acceptable, do not increase discrimination (if anything, the moderate justification in our experiment dampened discrimination). Third, extreme justifications not only increase discriminatory behavior but also shape individuals' reasoning. Subjects

who observe extreme justifications are more likely to include similarly extreme rationales among their own justifications. This effect is not mediated by a shift in perceptions of the social appropriateness of the statements — subjects do not update their beliefs about the social acceptability of extreme statements after being exposed to them. Instead, they adopt an extreme justification despite recognizing that using this justification is socially inappropriate.

Taken together, these findings represent a potentially novel dynamic mechanism of polarization: extreme language does not trigger backlash or deterrence but instead reinforces and spreads both discriminatory behavior and extreme rhetoric. This points to the possibility of a "vicious circle" whereby the presence of extreme justifications fuels greater discrimination and normalizes radical language, thus having the potential to intensify political polarization over time.

References

Barr, A., Lane, T., and Nosenzo, D. (2018). "On the social inappropriateness of discrimination". *Journal of Public Economics* 164, pp. 153–164.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). "Social proximity and the erosion of norm compliance". *Games and Economic Behavior* 132, pp. 59–72.

Bicchieri, C. and Xiao, E. (2009). "Do the right thing: but only if others do so". *Journal of Behavioral Decision Making* 22 (2), pp. 191–208.

Bisin, A. and Verdier, T. (2001). "The economics of cultural transmission and the dynamics of preferences". *Journal of Economic theory* 97 (2), pp. 298–319.

— (2011). "The economics of cultural transmission and socialization". *Handbook of social economics*. Vol. 1. Elsevier, pp. 339–416.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2025). "Inaccurate statistical discrimination: An identification problem". *Review of Economics and Statistics*, pp. 1–16.

Bohren, J. A., Hull, P., and Imas, A. (2022). *Systemic discrimination: Theory and measurement*. Tech. rep. w29820. National Bureau of Economic Research.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). "The dynamics of discrimination: Theory and evidence". *American Economic Review* 109 (10), pp. 3395–3436.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2024). "Cross-country trends in affective polarization". *Review of Economics and Statistics* 106 (2), pp. 557–565.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020). "From extreme to mainstream: The erosion of social norms". *American Economic Review* 110 (11), pp. 3522–3548.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., and Roth, C. (2023). "Justifying dissent". *The Quarterly Journal of Economics* 138 (3), pp. 1403–1451.

Carrell, S. E., Malmstrom, F. V., and West, J. E. (2008). "Peer effects in academic cheating". *Journal of Human Resources* 43 (1), pp. 173–207.

Charness, G. and Chen, Y. (2020). "Social identity, group behavior, and teams". *Annual Review of Economics* 12 (1), pp. 691–713.

Chen, D. L., Schonger, M., and Wickens, C. (2016). "oTree – An open-source platform for laboratory, online, and field experiments". *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.

Chen, M. K. and Rohla, R. (2018). "The effect of partisanship and political advertising on close family ties". *Science* 360 (6392), pp. 1020–1024.

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

Dimant, E. (2019). "Contagion of pro-and anti-social behavior among peers and the role of social proximity". *Journal of Economic Psychology* 73, pp. 66–88.

— (2024). "Hate trumps love: The impact of political polarization on social preferences". *Management Science* 70 (1), pp. 1–31.

Djourelova, M. (2023). "Persuasion through slanted language: Evidence from the media coverage of immigration". *American Economic Review* 113 (3), pp. 800–835.

Eyting, M., Buschinger, C., Hett, F., and Kessler, J. (2024). *Justifications*. Tech. rep. Retrieved from https://osf.io/9yb5v/.

Eyting, M. (2022). *Why do we discriminate? The role of motivated reasoning*. Tech. rep. 356. SAFE Working Paper.

Gächter, S., Gerhards, L., and Nosenzo, D. (2017). "The importance of peers for compliance with norms of fair sharing". *European Economic Review* 97, pp. 72–86.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). "The origins and consequences of affective polarization in the United States". *Annual Review of Political Science* 22 (1), pp. 129–146.

Iyengar, S., Sood, G., and Lelkes, Y. (2012). "Affect, not ideology: A social identity perspective on polarization". *Public Opinion Quarterly* 76 (3), pp. 405–431.

Iyengar, S. and Westwood, S. J. (2015). "Fear and loathing across party lines: New evidence on group polarization". *American Journal of Political Science* 59 (3), pp. 690–707.

Keizer, K., Lindenberg, S., and Steg, L. (2008). "The spreading of disorder". *Science* 322 (5908), pp. 1681–1685.

Kessler, J. B. (2017). "Announcements of support and public good provision". *American Economic Review* 107 (12), pp. 3760–3787.

Kranton, R., Pease, M., Sanders, S., and Huettel, S. (2020). "Deconstructing bias in social preferences reveals groupy and not-groupy behavior". *Proceedings of the National Academy of Sciences* 117 (35), pp. 21185–21193.

Krupka, E. L. and Weber, R. A. (2013). "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association* 11 (3), pp. 495–524.

Lane, T., Miller, L., and Rodriguez, I. (2024). "The normative permissiveness of political partyism". *European Economic Review* 162, p. 104661.

Shayo, M. (2020). "Social identity and economic policy". *Annual Review of Economics* 12 (1), pp. 355–389.

Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971). "Social categorization and intergroup behaviour". *European Journal of Social Psychology* 1 (2), pp. 149–178.

Wilson, J. Q. and Kelling, G. L. (1982). "Broken windows". *Atlantic Monthly* 249 (3), pp. 29–38.

Online Appendix

A Experiment

A.1 Experimental Design

This section provides sample screenshots of all relevant screens in the experiment.

A.1.1 Avatars

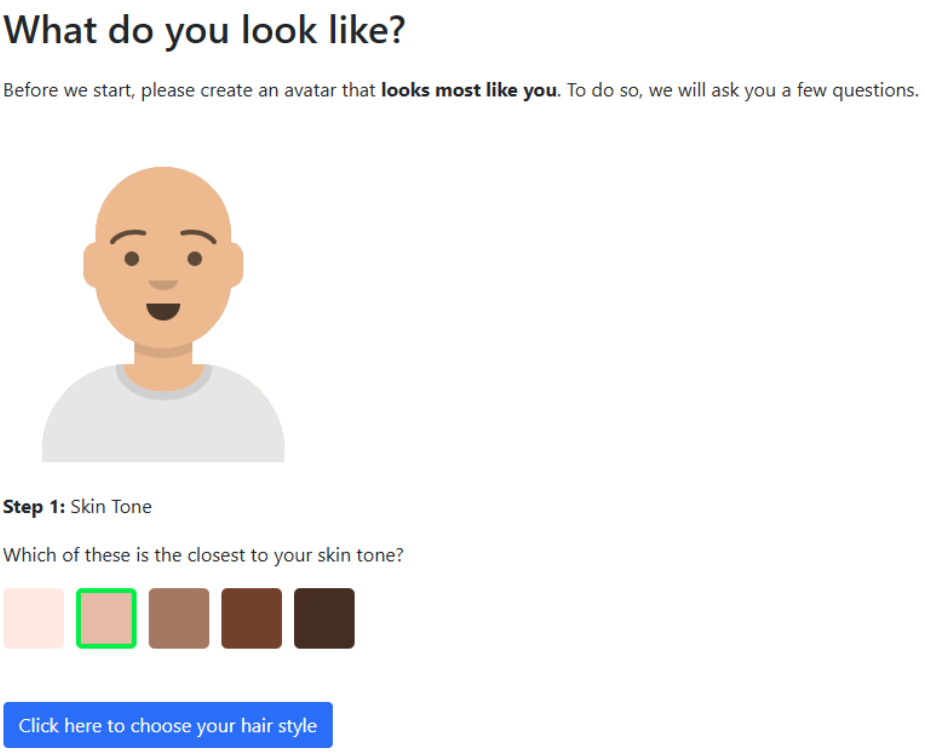


Figure A1: The figure shows the first step of the avatar creation, the selection of the skin color.

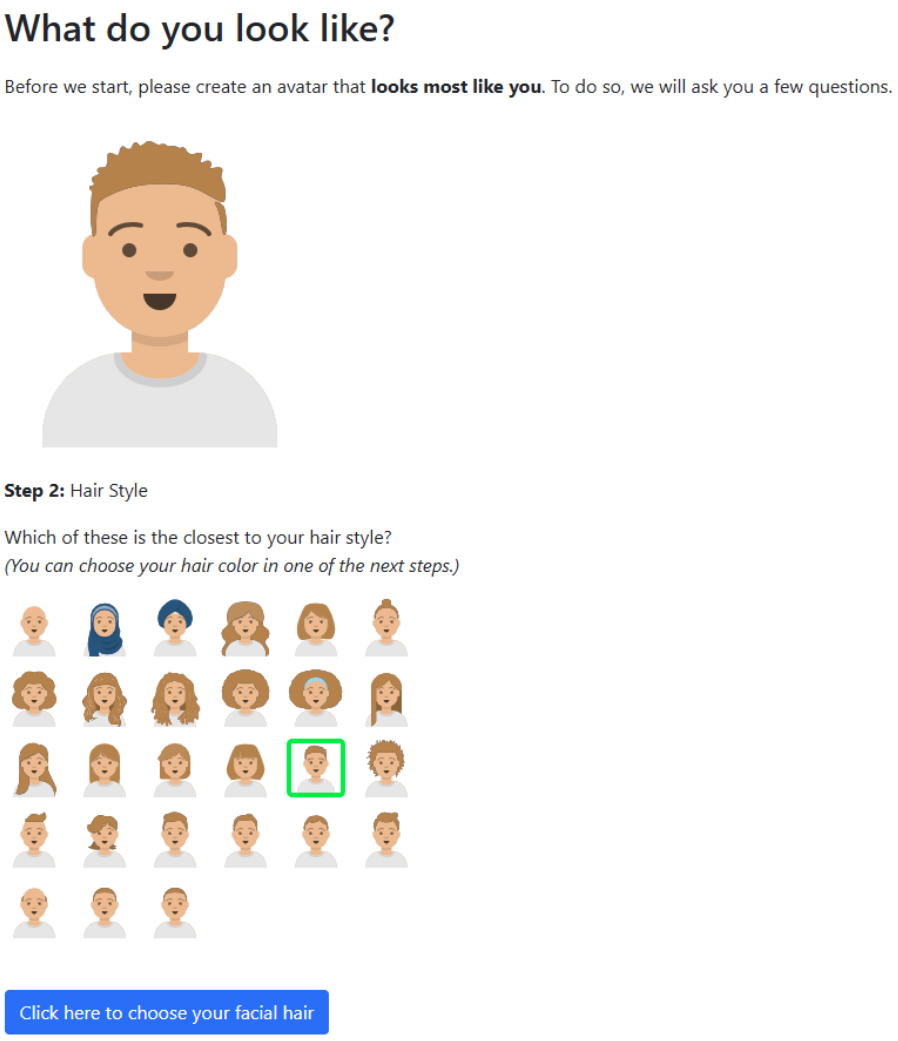


Figure A2: The figure shows the second step of the avatar creation, the selection of the hair style.

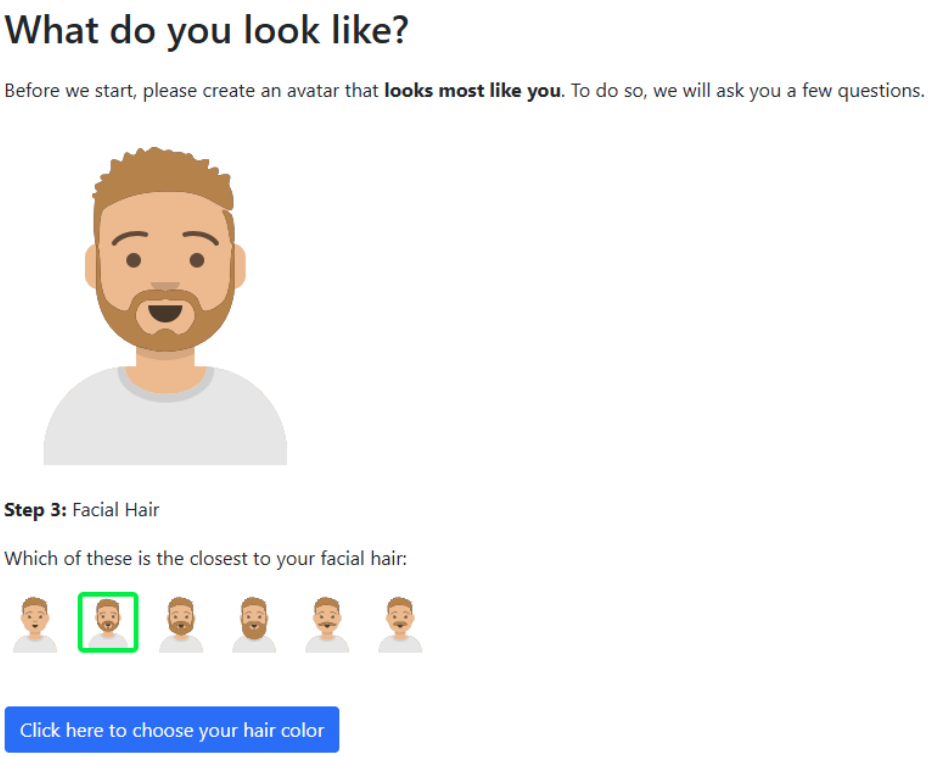


Figure A3: The figure shows the third step of the avatar creation, the selection of the beard style.

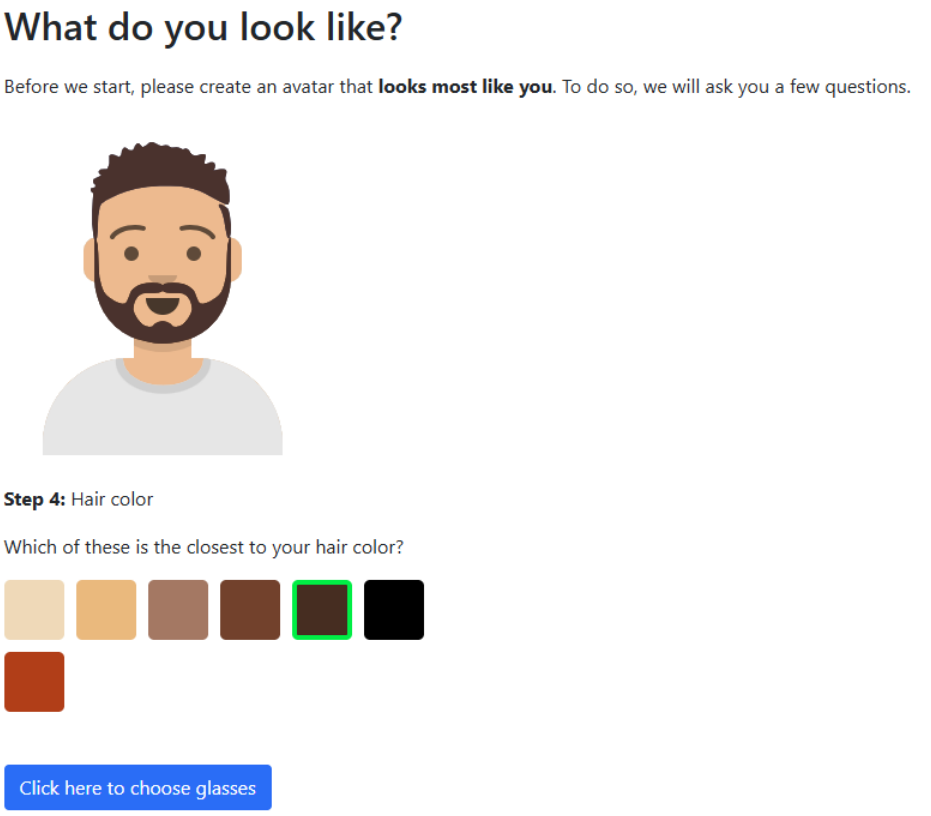


Figure A4: The figure shows the fourth step of the avatar creation, the selection of the hair (and beard) color.

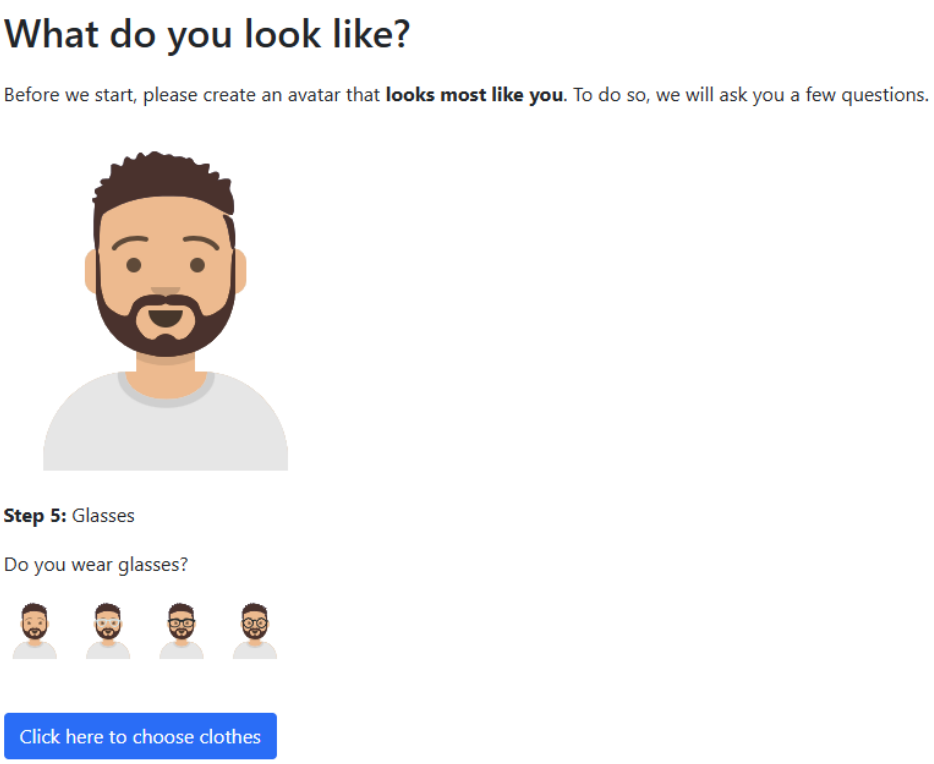


Figure A5: The figure shows the fifth step of the avatar creation, the selection of glasses.

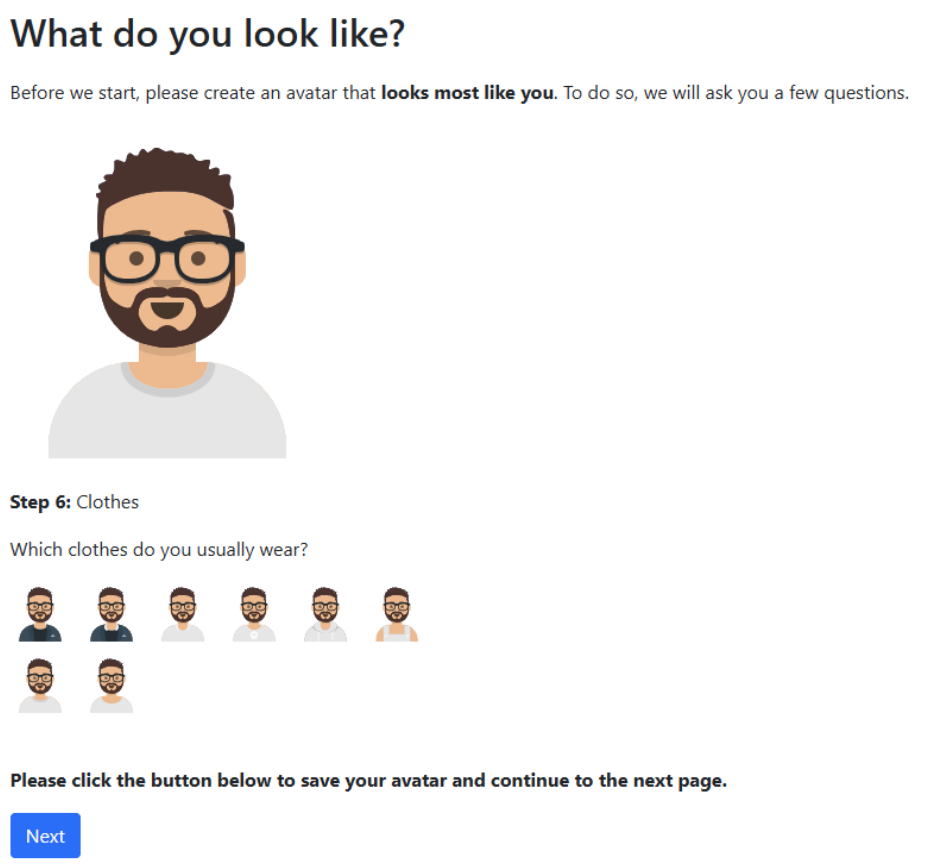


Figure A6: The figure shows the sixth step of the avatar creation, the selection of clothes.

A.1.2 Instructions

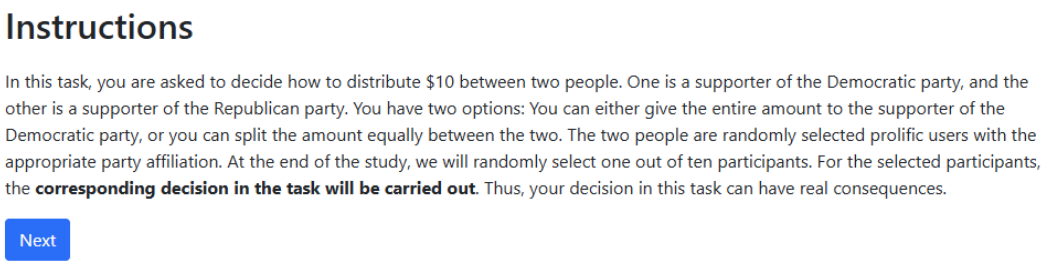


Figure A7: The figure shows the instructions of the main task.

A.1.3 Treatments

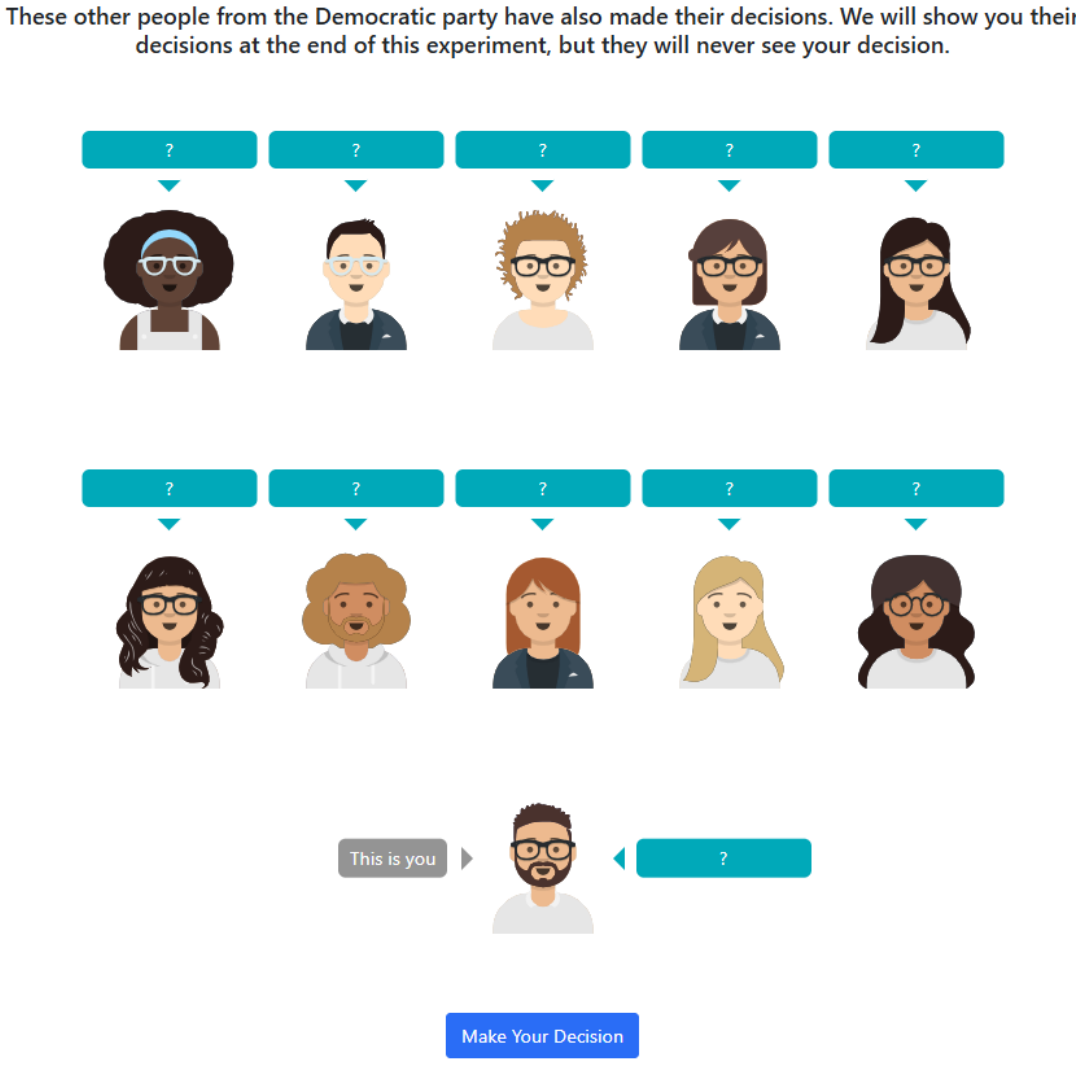


Figure A8: The figure shows what you see about peers in the "Baseline" treatment.

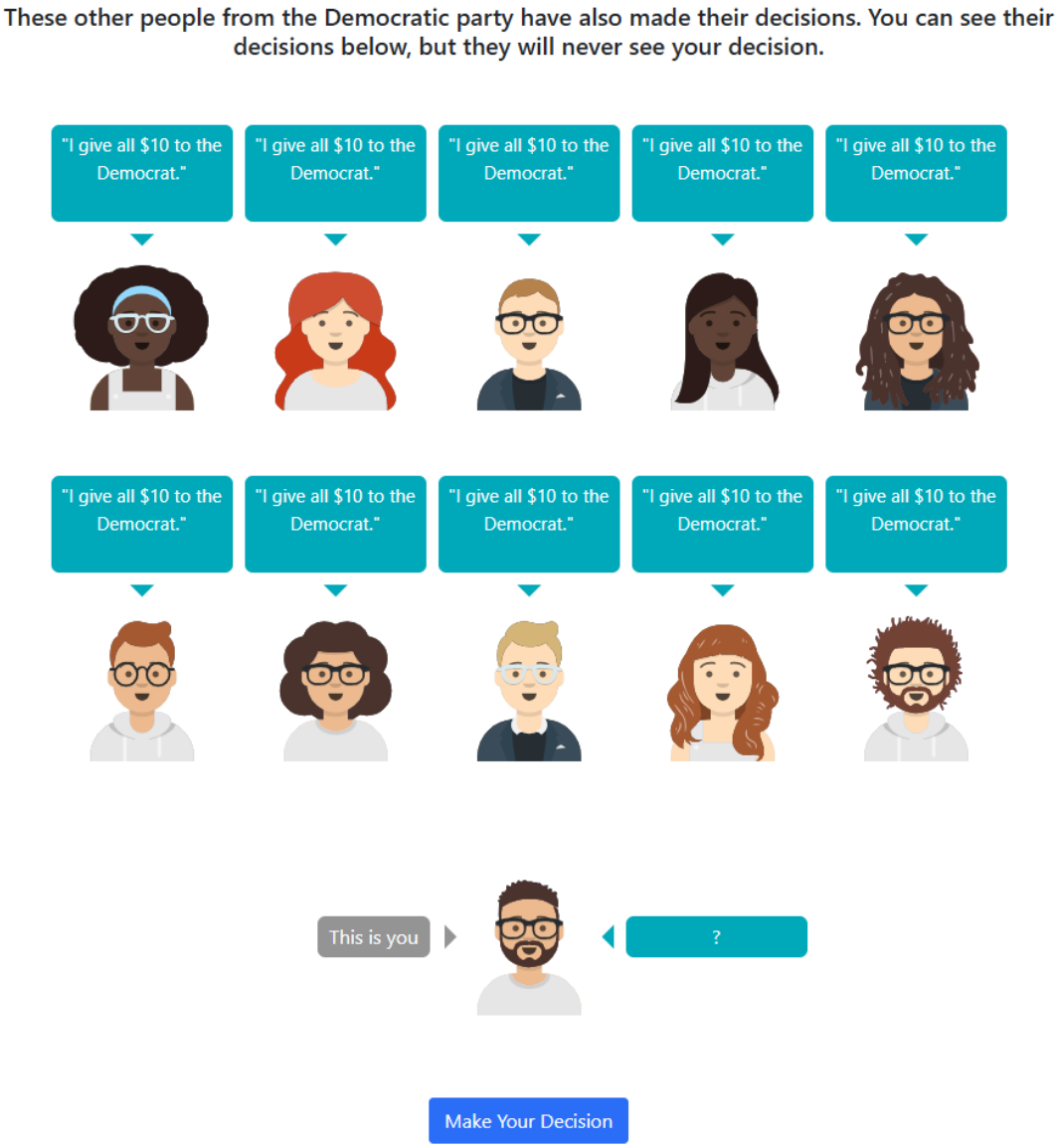


Figure A9: The figure shows an example of what subjects see about peers in the "No Justification" treatment.

Figure A10: The figure shows an example of what subjects see about peers in the "Moderate Justification" treatment.

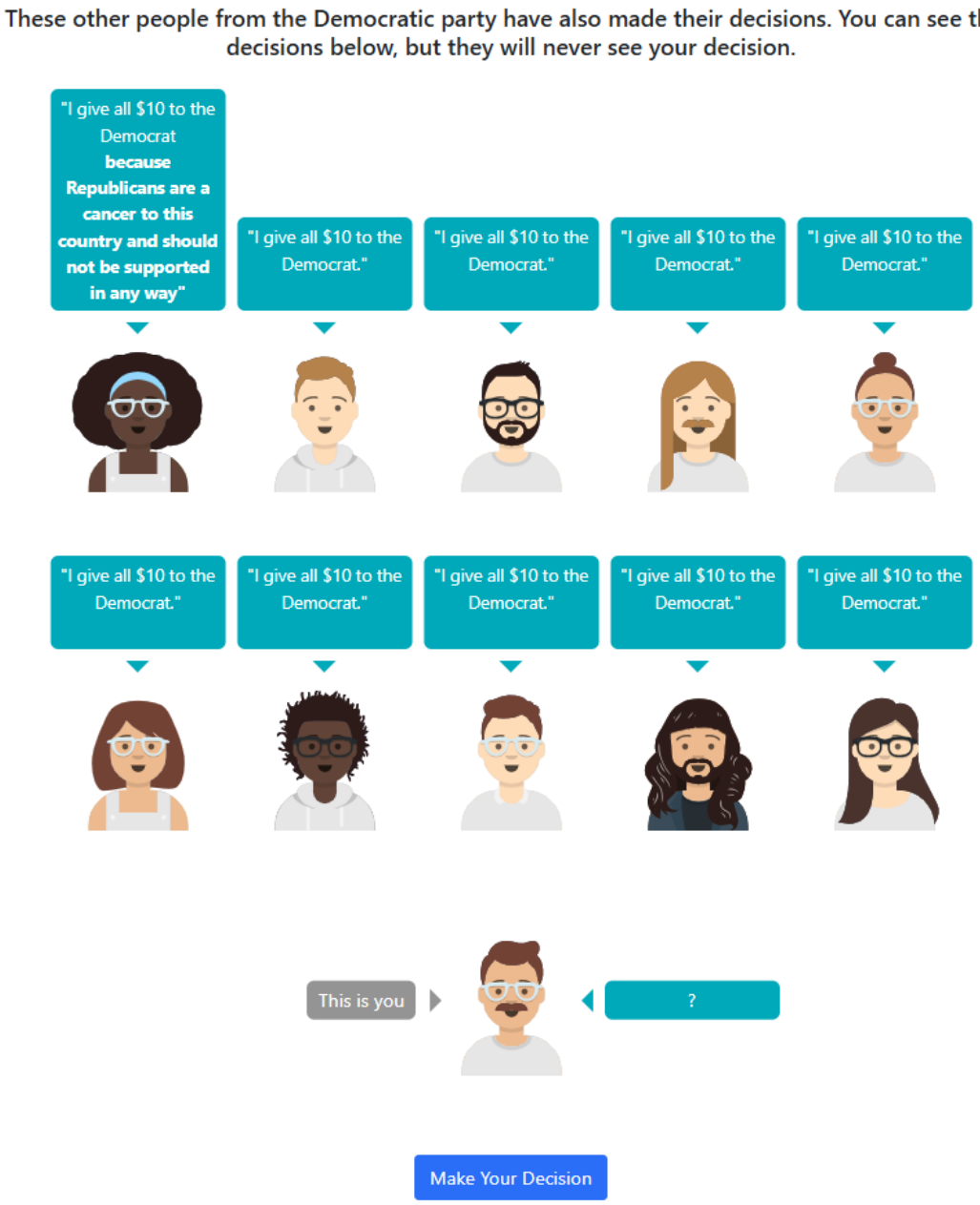


Figure A11: The figure shows an example of what subjects see about peers in the "Extreme Justification" treatment.

A.1.4 Main Decision Stage

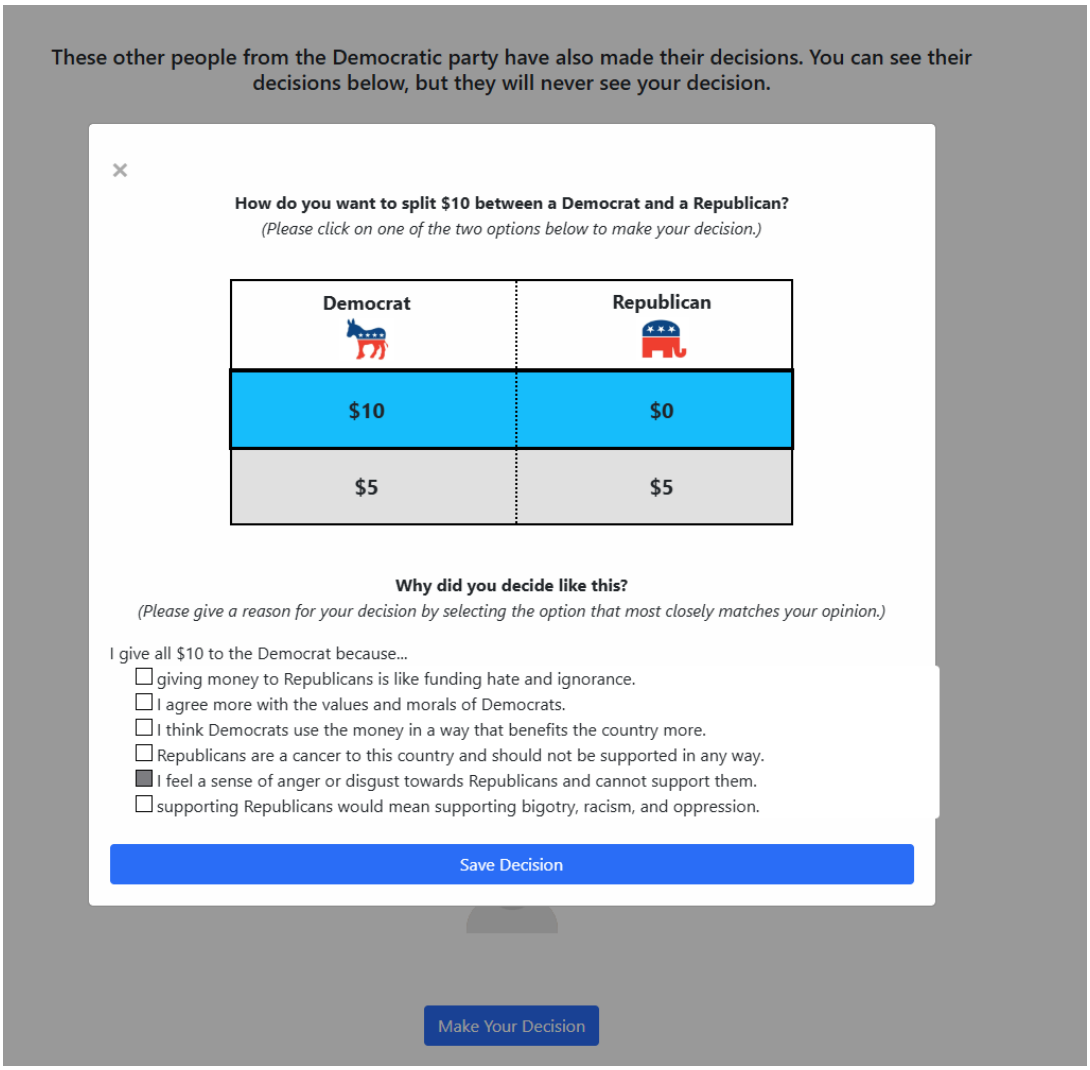


Figure A12: The figure shows the main decision from the perspective of a Democrat who chooses the unequal split. Subjects enter this as a pop up from the previous decision screen. This looks similar for all treatments. The six statements from which participants could choose depended on the decision (equal or unequal split), and the order of the statements was randomized.

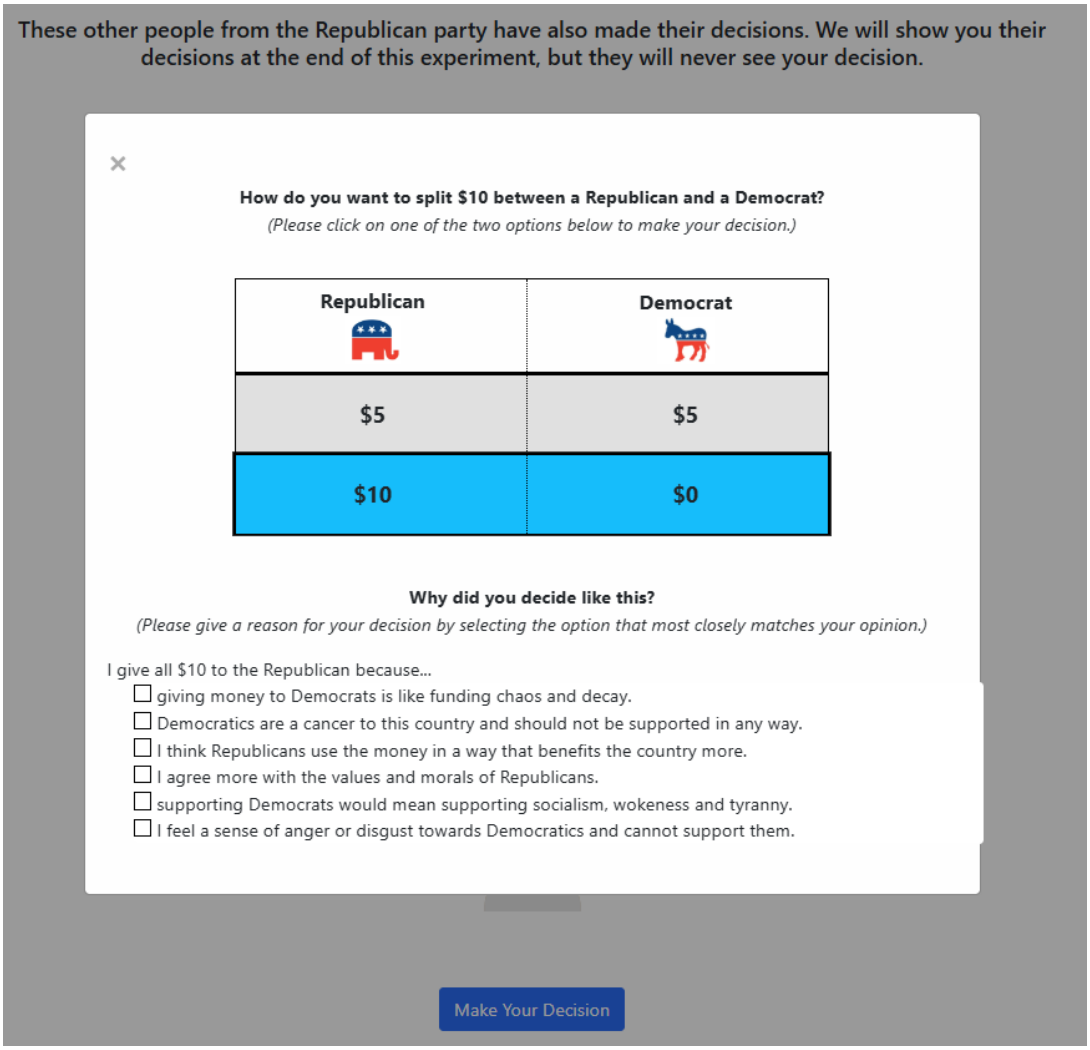


Figure A13: The figure shows the main decision from the perspective of a Republican who chooses the unequal split. Subjects enter this as a pop up from the previous decision screen. This looks similar for all treatments. The six statements from which participants could choose depended on the decision (equal or unequal split), and the order of the statements was randomized.

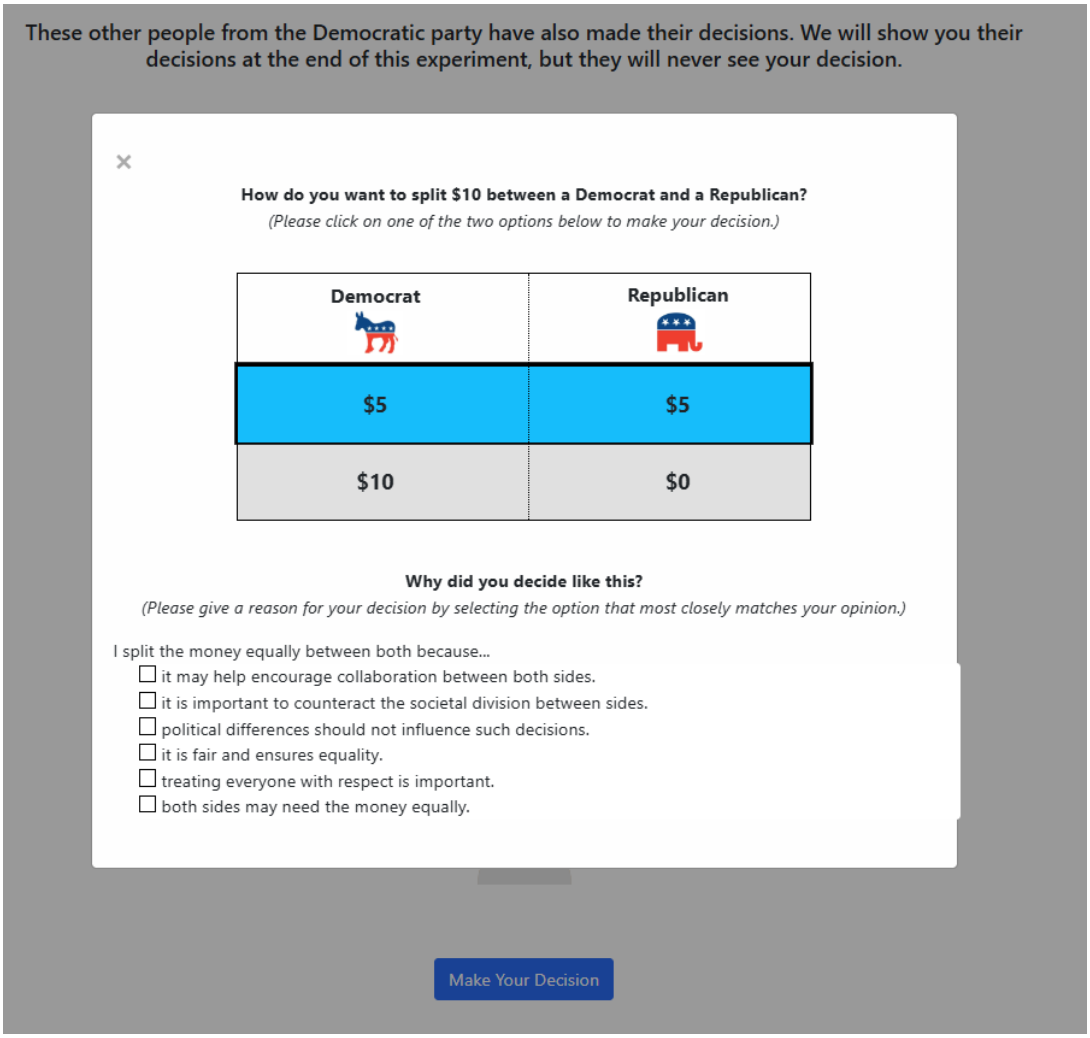


Figure A14: The figure shows the main decision from the perspective of a Democrat who chooses the equal split. Subjects enter this as a pop up from the previous decision screen. This looks similar for all treatments. The order of the statements was randomized. This screen looked similar for Republicans, except for the order of the parties.

Figure A15: The figure shows the last step of the main decision. Subjects see this overview after making their decision in the pop-up. The justification in their speech bubble is the one they previously selected in the pop-up window. They can either "Submit" or "Change" their decision. The "Change Decision" button reopens the pop-up window with the decision and justification they previously made.

A.1.5 Beliefs

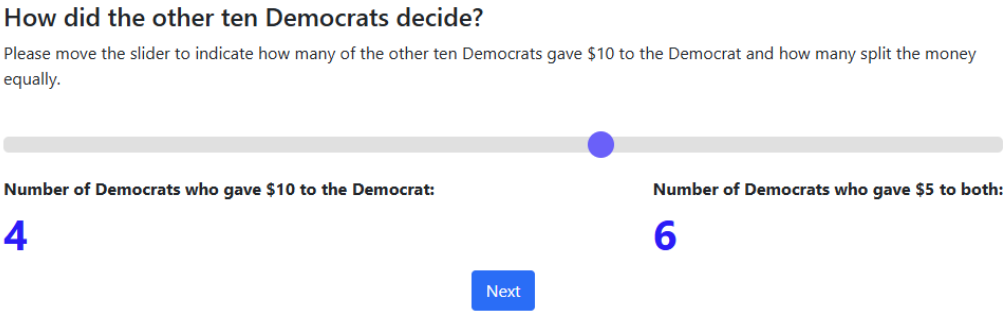


Figure A16: The figure shows the belief elicitation question about the decisions of the 10 peers. In the "Baseline" treatment we incentivize the decision by adding a sentence with the incentive. Democrats saw: "If you correctly guess the decisions of the other ten Democrats you receive $0.50." Republicans saw the same sentence with the word Democrats replaced with the word Republicans.

A.1.6 Further Justifications

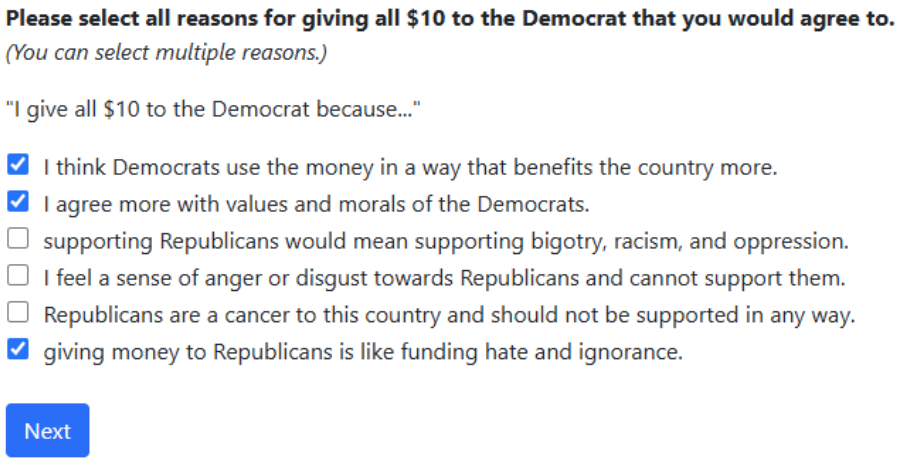


Figure A17: The figure shows the question inviting subjects to select all of the justifications with which they agree. The same six reasons as in the pop-up on the the main decision page were displayed here. The six statements shown to subjects depended on whether the subject choose to provide an equal or unequal split. The order of the statements was randomized.

A.1.7 Social Appropriateness (Wave 2)

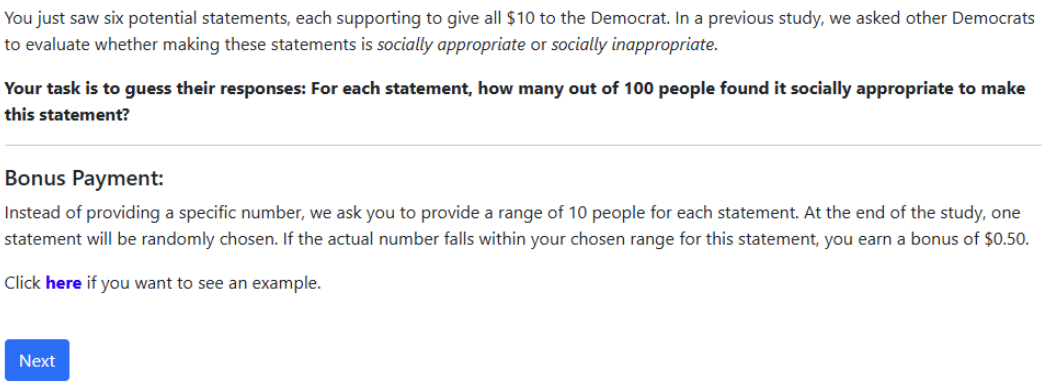


Figure A18: The figure shows the instructions for the elicitation of beliefs about the social appropriateness of the statements. Only participants in wave 2 completed this task. We asked subjects to rate each of the six justifications available to subjects who chose an unequal split.

Figure A19: The figure shows the elicitation of the belief about the social appropriateness of the statements. Only participants in wave 2 completed this task. The figure shows a case where the subject has made a guess for the first two statements and not yet made a guess for the other four. The order of the statements was randomized.

A.1.8 Debriefing Survey and Attention Check

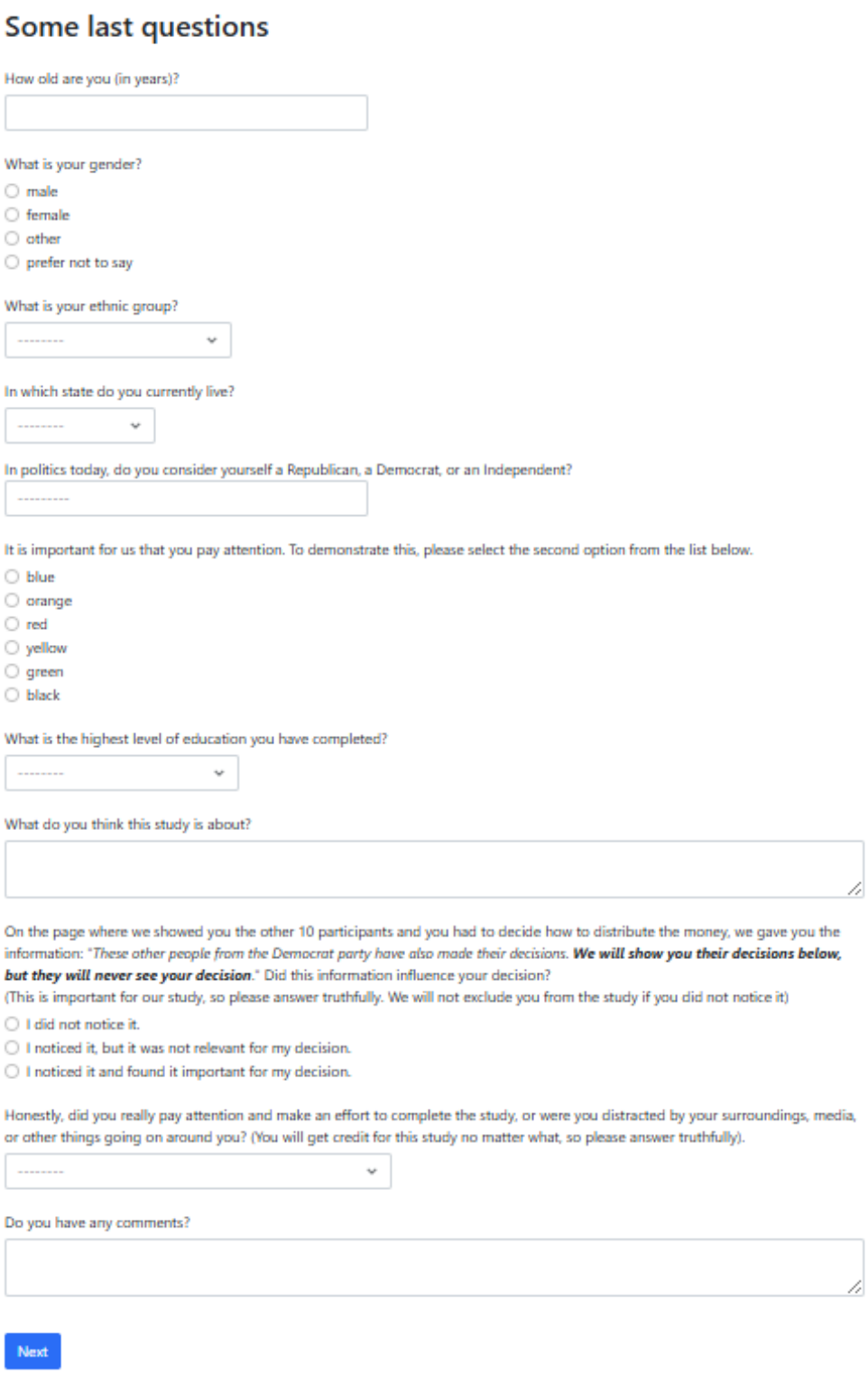


Figure A20: The figure shows the debriefing survey, which collects demographic data to confirm the political affiliation of the subject and checks participants' attentiveness. The third-to-last and second-to-last questions were added in wave 2.

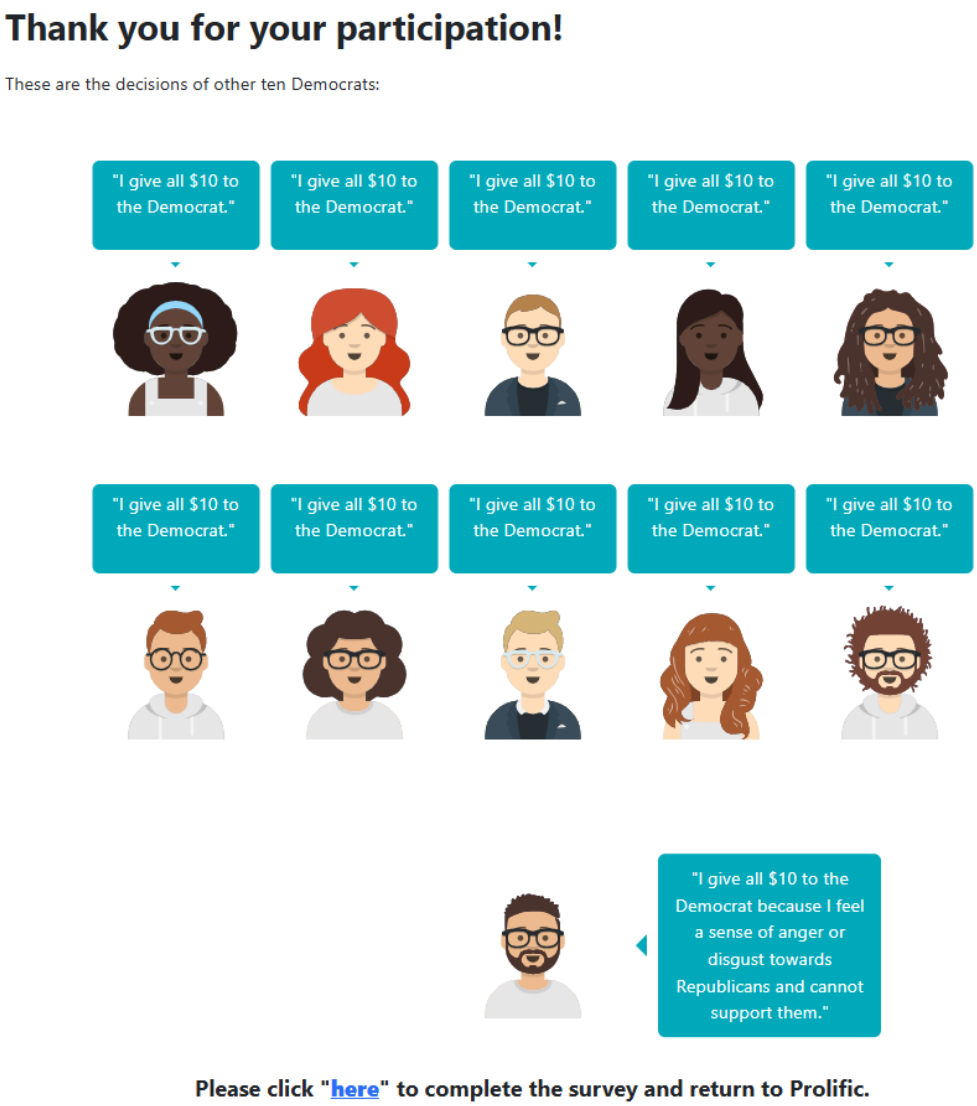


Figure A21: The figure shows the last page of the study when subjects again see their ten peers and the decisions they made. We included this "Debriefing" page because we told participants in the "Baseline" treatment that they would learn about their peers' decisions at the end of the study.

A.2 Comparison of Waves

As noted in the main text, the two waves that we ran differ slightly in some details of the experimental design.

First, for wave 1, peer information was elicited in a small pre-study run a few days before our main wave 1 data collection. For wave 2, peer information came from subjects in wave 1.

Second, in wave 2 we added a task to elicit beliefs about the social appropriateness of the justifications for discrimination, as shown in Figure A19. We added a few additional questions to the debriefing survey at the end of the study. All other parts of the study, particularly the treatments and main decisions, stayed the same. The additional task and survey questions increased the time subjects needed to complete the survey from 5 to 8 minutes on average and thus we raised the participation fee to keep the hourly wage approximately the same across waves.

Third, we slightly changed how we selected peers across waves. In wave 1, the peers were drawn randomly from all members of the subject's political party (except for the first avatar on the top left). In wave 2 we stratified how many of the 10 peers discriminated. This was done to get more data for cases where very few or very many of the ten peers discriminated, which came up rarely in wave 1.

Fourth, the first avatar — that is always selected to be a peer who discriminated against the outgroup — was selected somewhat differently between the waves. In wave 1, we randomly drew one avatar from a pool of possible first avatars who discriminated and selected either the "Cancer" *or* "Values" statement as a justification with which they agreed. For the "Extreme Justification" and "Moderate Justification" treatments, we then restricted subjects to be shown an avatar that selected the corresponding statement. In wave 2, we changed our selection rule slightly so the first avatar was randomly selected from a pool of avatars who discriminated and selected both the "Cancer" *and* "Values" statement as a justification with which they agreed. This ensured that the distribution of avatars subjects saw was the same across all treatments in wave 2.

Finally, in wave 1, we had introduced slight variation in instructions within the Moderate Justification and Extreme Justification treatments. Some subjects were told: "[other participants] will never see your decision" while other subjects were told: "In one of the following sessions your avatar and decision will be at the top left, and thereby prominently shown to another participant". This variation was meant to induce social image concerns. However, results did not differ systematically by instructions, presumably because the treatment was too weak. Our regression results control for which instructions subjects received in wave 1, and we dropped this variation in wave 2.

All design features — and adjustments — were pre-registered under https://osf.io/j493t (wave 1) and https://osf.io/yuaej (wave 2).

A.3 Validation Study

As part of our experiment, we conducted a separate "Validation Study" in which a different group of subjects was asked to assess the extremity of the six statements. In this study, 130 each of Democrats, Republicans, and Independents were provided with information about the task from the main study (see Figure A22) before they rated the extremity and the social appropriateness of the six justification statements that could be chosen by subjects who discriminated.

The validation was conducted from the perspective of a Democrat for Democrats and a Republican for Republicans. For Independents, we randomly drew the framing of the evaluation task as either from the perspective of a Democrat or a Republican. We excluded participants for whom the prolific information on their political affiliation did not match their self-reported party preferences from a survey at the end of the validation study (as shown in Figure A25).

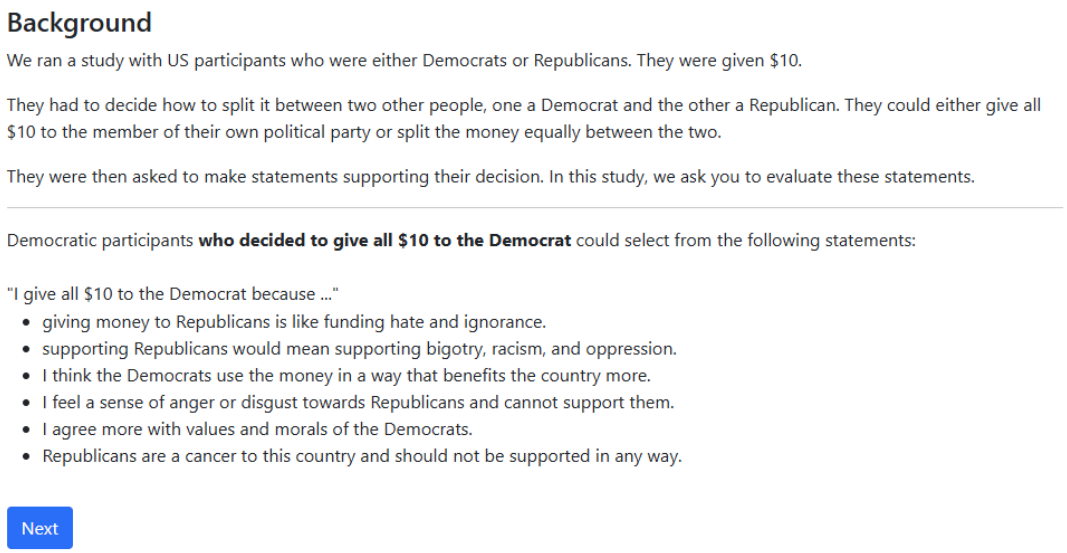


Figure A22: The figure shows the information subjects received about the decision from our main study before evaluating the statements.

In the first task — as shown in Figure A23 — participants assess each statement for how extreme they found it to be as a reason to favor the ingroup member in the allocation decision. This aims to verify the extremity level of our statements.

Figure A24 shows the second validation task. Here, we asked subjects to indicate whether they find it socially appropriate or inappropriate to say each statement. In addition to representing another perspective of perceived extremity of the statements to validate them, this task was also used to incentivize our elicitation of social appropriateness beliefs in wave 2 of our main study.

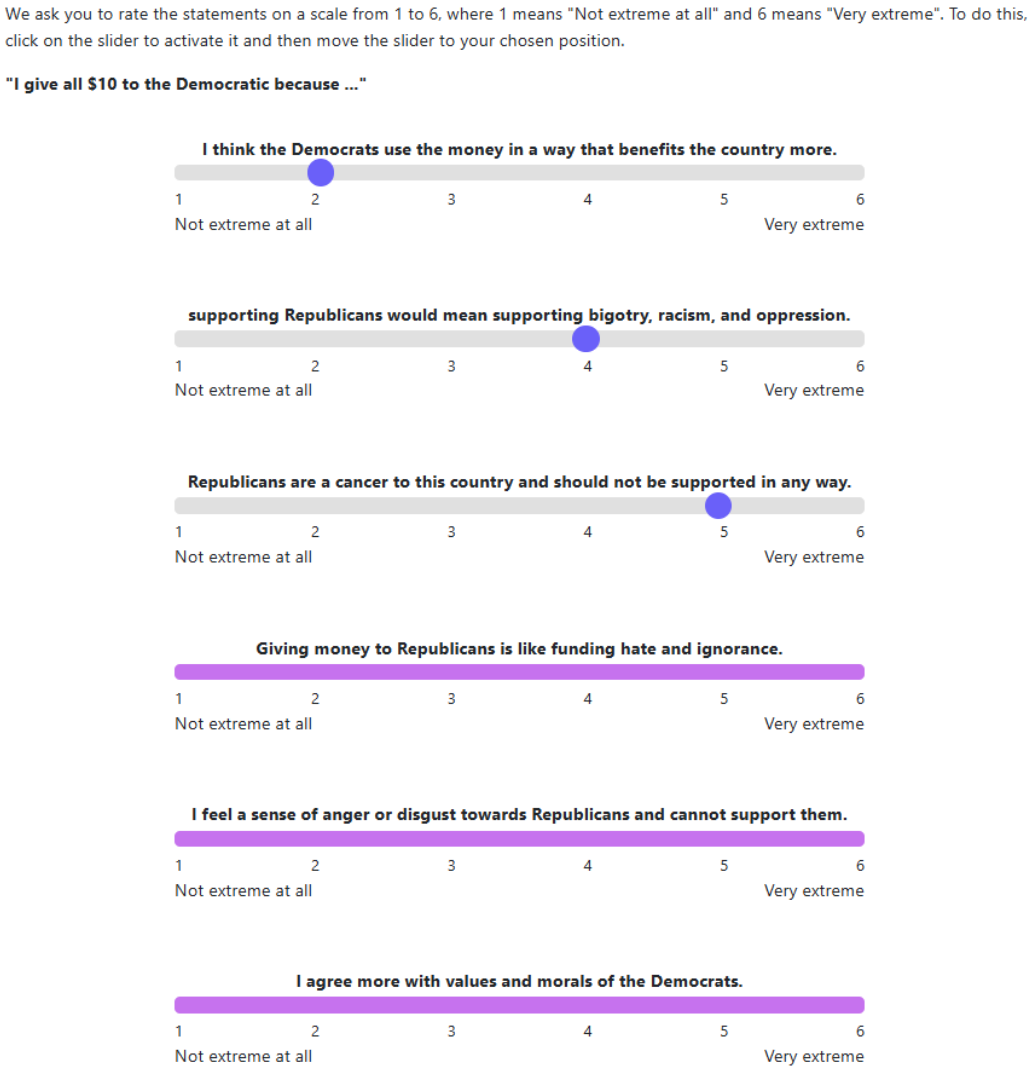


Figure A23: The figure shows the first evaluation task from the validation study, the evaluation of the statements' extremity. The order of the statements was randomized.

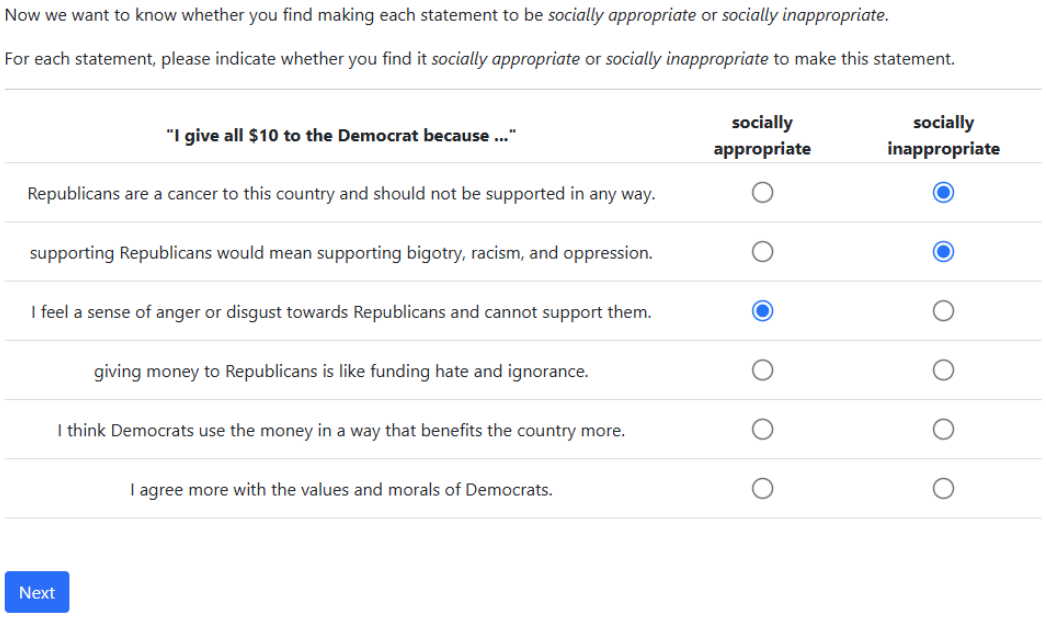


Figure A24: The figure shows the second evaluation task from the validation study, the evaluation of the statements' social appropriateness. The order of the statements was randomized.

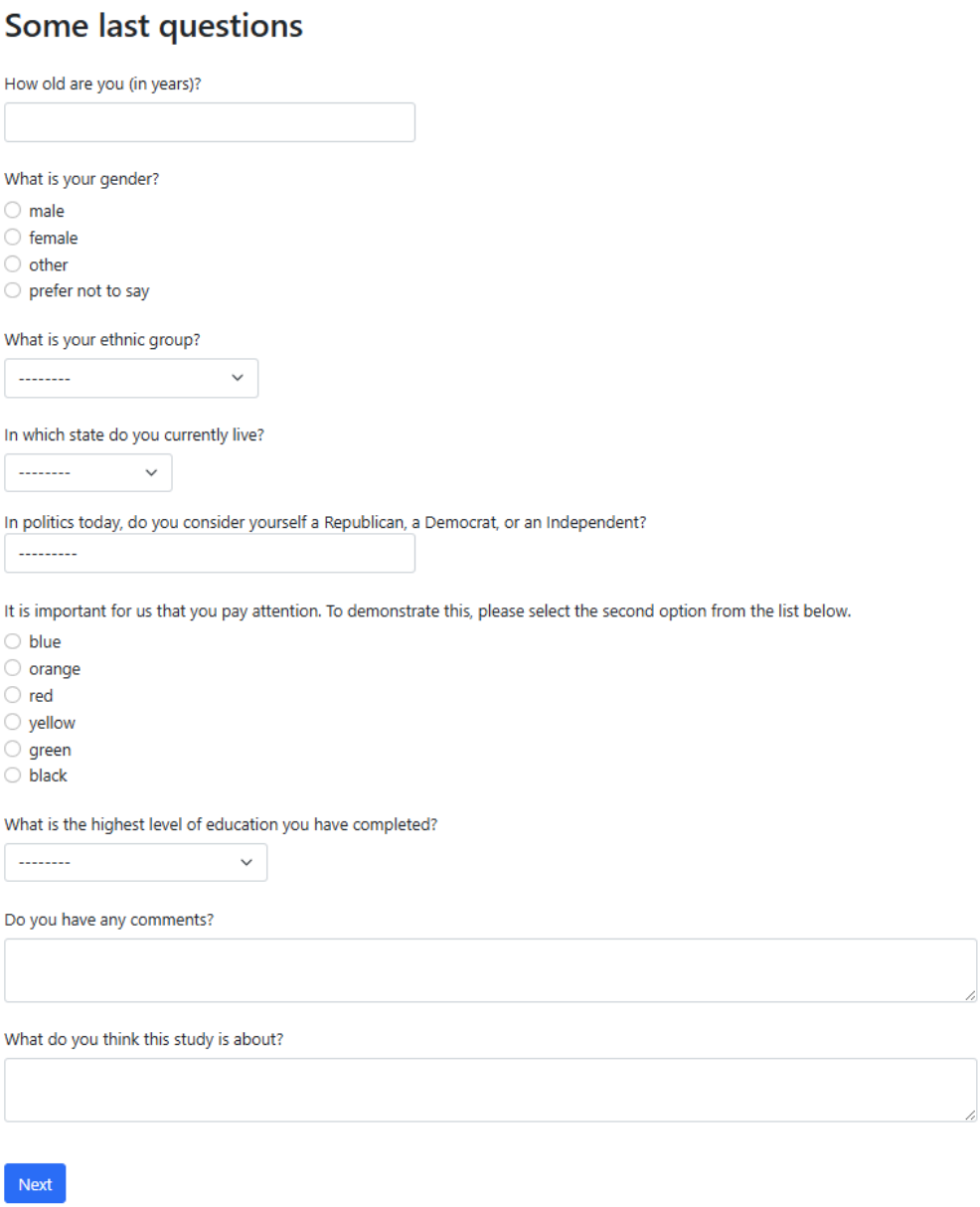


Figure A25: The figure shows the survey at the end of the validation study.

B Further Analyses

This section provides details on additional analysis that is discussed in the main body of the paper.

B.1 Validation Study: Extremity of Statements

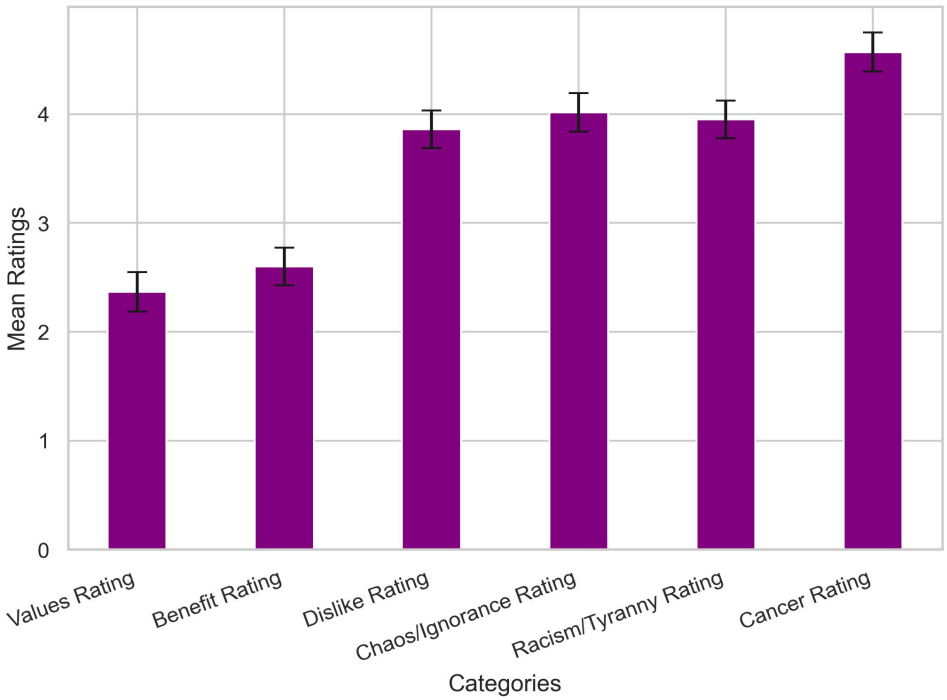
Table B1: Overview of Justification Statements

	Label	Statement	Extremity
Democrats	Values	"I agree more with the values and morals of Democrats."	moderate
	Benefit	"I think Democrats use the money in a way that benefits the country more."	moderate
	Dislike	"I feel a sense of anger and disgust towards Republicans and cannot support them."	intermediate
	Ignorance	"Giving money to Republicans is like funding hate and ignorance."	intermediate
	Racism	"Supporting Republicans would mean supporting bigotry, racism, and oppression."	intermediate
	Cancer	"Republicans are a cancer to this country and should not be supported in any way."	extreme
Republicans	Values	"I agree more with the values and morals of Republicans."	moderate
	Benefit	"I think Republicans use the money in a way that benefits the country more."	moderate
	Dislike	"I feel a sense of anger and disgust towards Democrats and cannot support them."	intermediate
	Chaos	"Giving money to Democrats is like funding chaos and decay."	intermediate
	Tyranny	"Supporting Democrats would mean supporting socialism, wokeness and tyranny."	intermediate
	Cancer	"Democrats are a cancer to this country and should not be supported in any way."	extreme

Notes: The table presents an overview of the justifications subjects could chose when they selected the unequal split and its label of extremity as classified based on the responses in the "Validation Study."

As indicated in Figure B1, "Cancer" is clearly perceived the most extreme statement among the six options. "Cancer" is rated as more extreme than the other five justifications. The difference is highly significant in the pooled version ($p < 0.001$) and in all cases at least weakly significant when looking at each subgroup separately ($p < 0.1$). "Values" — our statement in the "Moderate Justification" treatment — is perceived as the least extreme. While the difference between "Values" and "Benefit" is relatively small and only weakly significant in the pooled version ($p < 0.1$), all other justifications are assessed as clearly and highly statistically significantly more extreme ($p < 0.001$). A similar pattern can be observed in the three political subgroups separately.

Figure B1: Extremity Ratings



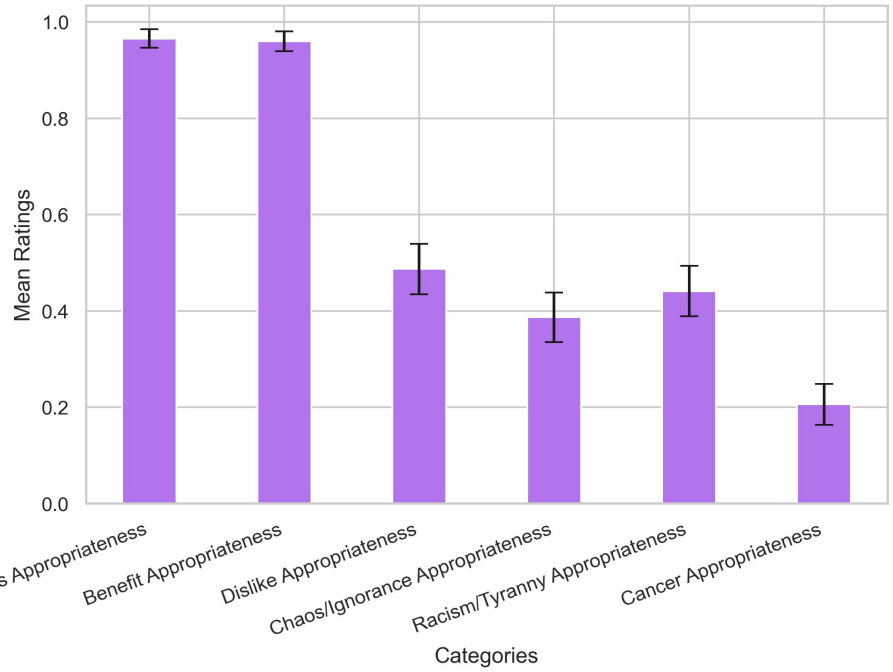
Notes: The figure shows the mean rating of the extremity of the statements indicated on the Likert scale as shown in Figure A23.

Thus, we can conclude that "Cancer" is indeed an extreme justification and "Values" is indeed a moderate one. In addition, our categorization of the six justifications into three groups "extreme," "intermediate," and "moderate" for additional analysis (e.g., for Appendix Figure B7) is supported by these results.

Note that in our analyses where we pool Democrats and Republicans, we combine the statements "Tyranny" and"Racism" as well as "Chaos" and "Ignorance". We do this as the structure of these statements is similar for each party. Our validation study (see A.3) further supports that they are perceived as approximately equally extreme.

Results on social appropriateness show a very similar picture to the extremity ratings analyzed above. "Values" and "Benefit" are clearly perceived as most socially appropriate justifications. Remarkably, both have mean ratings very close to 1, suggesting that almost all participants think that it is socially appropriate (rather than inappropriate) to use these statements to justify ingroup favoritism. "Cancer," on the other is clearly assessed as the most socially inappropriate reason. About 80% of participants say that this is socially inappropriate to make such a statement.

Figure B2: Social Appropriateness



Notes: The figure shows the mean rating of the social appropriateness of the statements obtained from the binary decision as shown in Figure A24.

B.2 Main Sample: Descriptive Statistics

Table B2: Summary Statistics

Variable	Pooled		Wave 1		Wave 2	
	Mean	Std	Mean	Std	Mean	Std
Democrat	0.50	0.50	0.50	0.50	0.50	0.50
Age	41.93	19.04	40.49	14.22	43.19	22.33
Male	0.49	0.50	0.49	0.50	0.48	0.50
Race						
White	0.67	0.47	0.61	0.49	0.72	0.45
Black	0.22	0.41	0.26	0.44	0.18	0.39
Hispanic	0.05	0.22	0.05	0.23	0.05	0.21
Asian	0.05	0.22	0.06	0.24	0.04	0.20
Other	0.02	0.12	0.02	0.14	0.01	0.11
Education						
No High School	0.01	0.08	0.01	0.08	0.01	0.07
High School	0.12	0.32	0.14	0.34	0.10	0.30
Began college	0.15	0.36	0.16	0.36	0.14	0.35
Associate	0.09	0.28	0.09	0.29	0.08	0.27
Bachelor	0.39	0.49	0.40	0.49	0.38	0.49
Master	0.21	0.41	0.17	0.38	0.24	0.43
Doctoral	0.04	0.20	0.03	0.17	0.05	0.21
Observations	2471		1150		1321	

Note: This table presents descriptive statistics for both waves and pooled. Summary statistics are shown for our final sample after excluding participants whose self-reported partisanship did not match the Prolific information and those who failed the attention check.

Table B3: Randomization Check Treatments

Variables	Baseline	No Justification	Extreme Justification	Moderate Justification	p-value
Age	40.17	41.88	42.74	42.07	0.1934
Democrat	0.50	0.50	0.51	0.49	0.9703
Gender					0.8205
Female	0.53	0.50	0.51	0.49	
Male	0.47	0.50	0.48	0.50	
Other	0.01	0.01	0.01	0.01	
Education					0.9629
Associate	0.10	0.08	0.08	0.09	
Bachelor	0.39	0.38	0.40	0.39	
Began College	0.15	0.15	0.14	0.15	
Doctoral	0.04	0.05	0.04	0.04	
High School	0.13	0.13	0.12	0.11	
Master	0.19	0.20	0.21	0.22	
No High School	0.01	0.00	0.01	0.01	
Other	0.01	0.00	0.00	0.00	
Ethnicity					0.8011
Asian	0.05	0.05	0.05	0.06	
Black or African American	0.23	0.22	0.20	0.22	
Hispanic or Latin	0.06	0.05	0.04	0.05	
White	0.64	0.66	0.69	0.66	
other / prefer not to answer	0.02	0.01	0.02	0.02	
Observations	**387**	**570**	**756**	**758**	

Notes: This table compares mean values of demographics between conditions. p-values are from a one-wave ANOVA test for mean differences between conditions for Age and Democrat and from a Chi-square test for differences in the frequency of the categories between conditions for Gender, Education, and Ethnicity.

B.3 Effect of Justifications on Discrimination

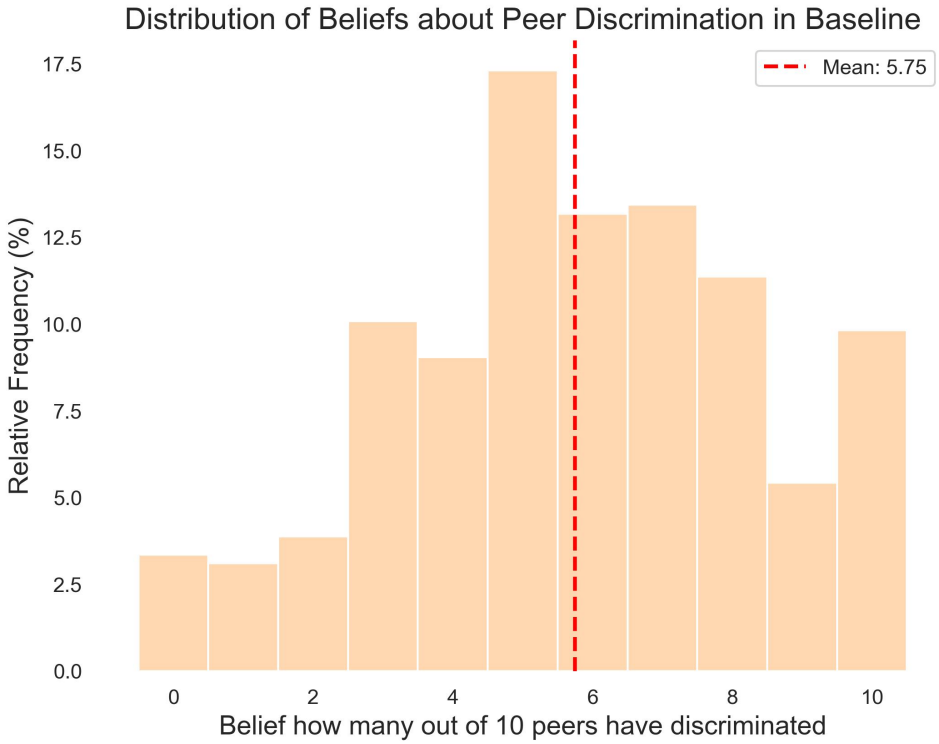
Table B4: The Effects of Justifications on Outgroup Discrimination by Political Party

Variables	Pooled	Republicans	Democrats
	(1)	(2)	(3)
Extreme Justification	0.051*	0.054	0.042
	(0.028)	(0.041)	(0.039)
Moderate Justification	-0.032	-0.066	0.000
	(0.028)	(0.040)	(0.039)
No Justification Mean	0.567	0.491	0.643
Extreme Justification – Moderate Justification			
Difference	0.083	0.120	0.041
p-value	0.001***	0.001***	0.243
Wave × No. of Peers Dummies	X	X	X
Instructions × No. of Peers Dummies	X	X	X
Observations	2084	1043	1041

Notes: This table presents regression results on discrimination controlling for dummies for wave and instructions in wave 1, each interacted with dummies for the number of peers who discriminated. The "No Justification" treatment is the excluded group. Column (1) shows results for the main sample in the "No Justification," "Moderate Justification," and "Extreme Justification" treatments. Column (2) shows results for Republicans and column (3) shows results for Democrats. *p< 0.1, **p< 0.05, ***p< 0.01.

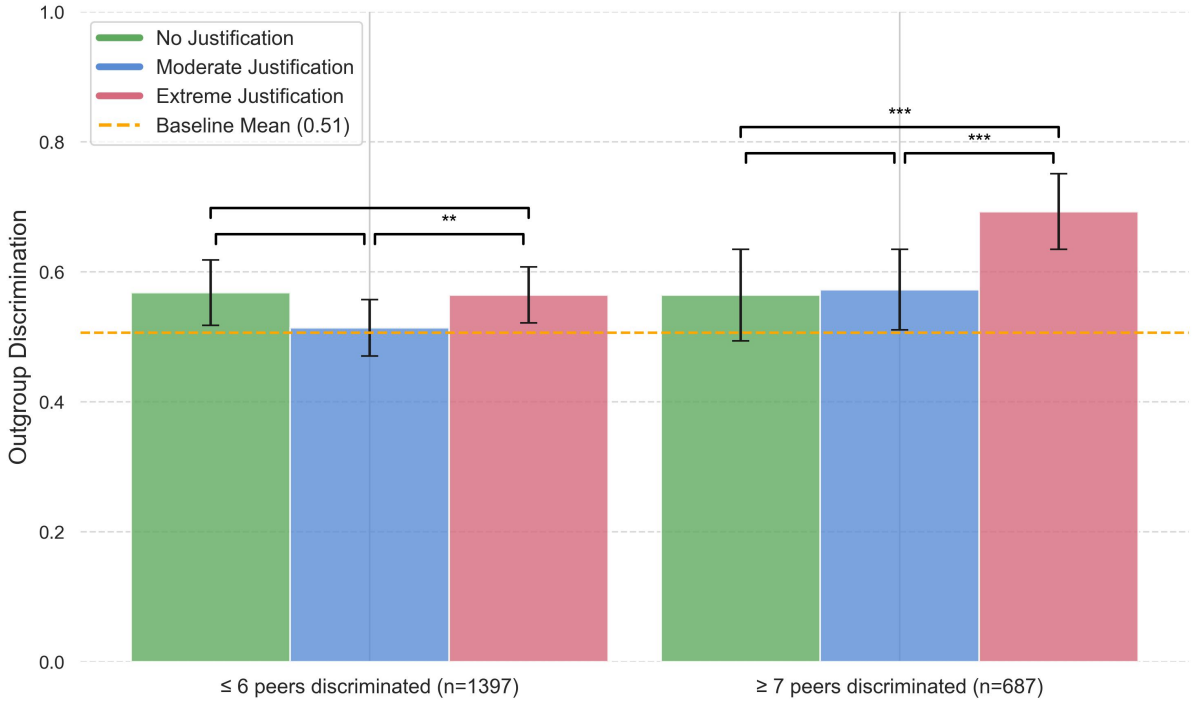
B.4 The Role of Peers' Behavior for the Effect of Justifications

Figure B3: Prior Beliefs about Number of Discriminating Peers



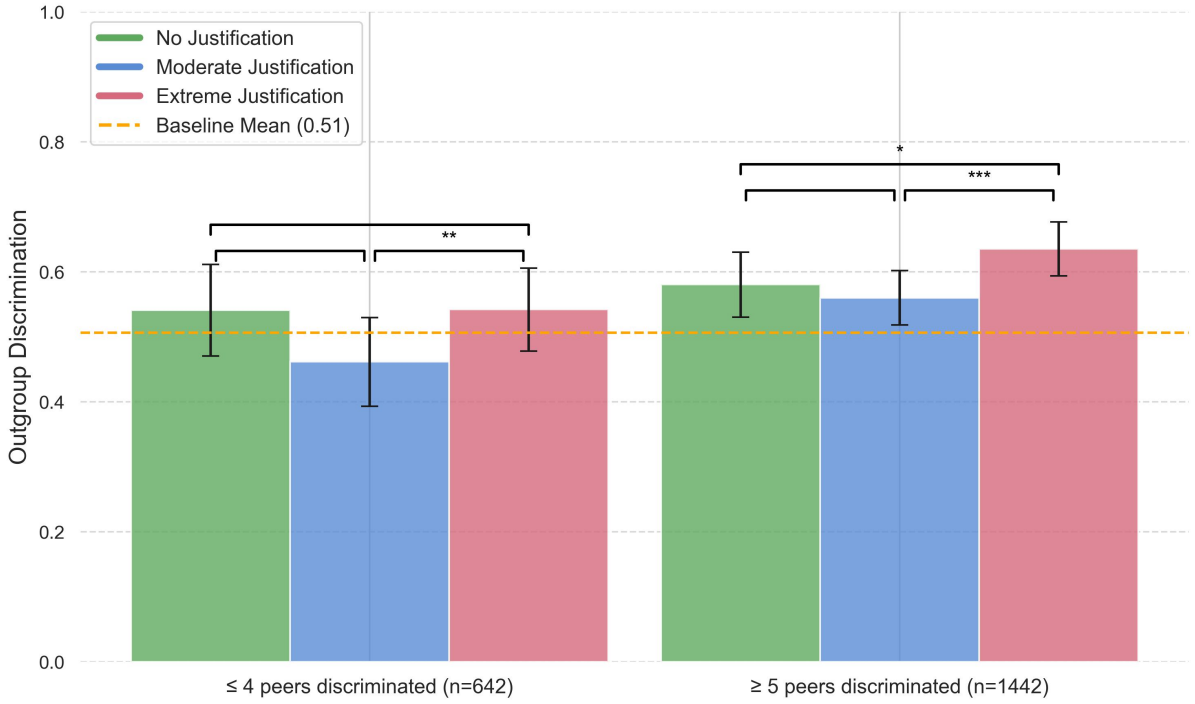
Notes: The figure shows the belief distribution in the "Baseline" treatment about how many of the 10 peers discriminated. This belief elicitation was incentivized as explained in Section 2.

Figure B4: The Effects of Justifications on Discriminatory Behavior across Others' Behavior, Alternative Split with ≤ 6 and ≥ 7



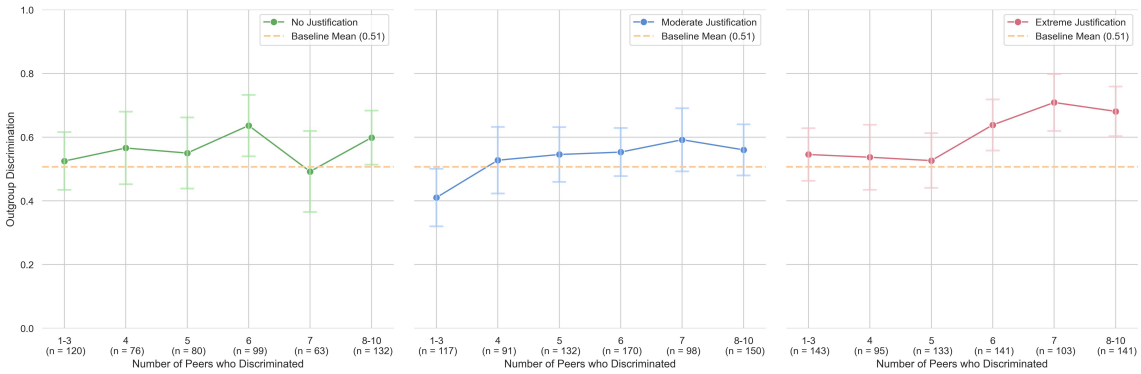
Notes: "Outgroup Discrimination" is a binary variable equal to 1 if the unequal split was chosen. The green, blue, and red bars show the likelihood of discrimination across the "No Justification," "Moderate Justification," and "Extreme Justification" treatments, respectively. The dashed yellow line shows the level of discrimination in the "Baseline" treatment. The three bars on the left include the subset of subjects who were randomly chosen to be shown ≤ 6 peers discriminating and the three bars on the right show results for subjects who were randomly chosen to be shown ≥ 7 peers discriminating. Significance stars are based on regression specifications that include the same set of controls in Table 1. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01.

Figure B5: The Effects of Justifications on Discriminatory Behavior across Others' Behavior, Alternative Split with ≤ 4 and ≥ 5



Notes: "Outgroup Discrimination" is a binary variable equal to 1 if the unequal split was chosen. The green, blue, and red bars show the likelihood of discrimination across the "No Justification," "Moderate Justification," and "Extreme Justification" treatments, respectively. The dashed yellow line shows the level of discrimination in the "Baseline" treatment. The three bars on the left include the subset of subjects who were randomly chosen to be shown ≤ 4 peers discriminating and the three bars on the right show results for subjects who were randomly chosen to be shown ≥ 5 peers discriminating. Significance stars are based on regression specifications that include the same set of controls in Table 1. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01.

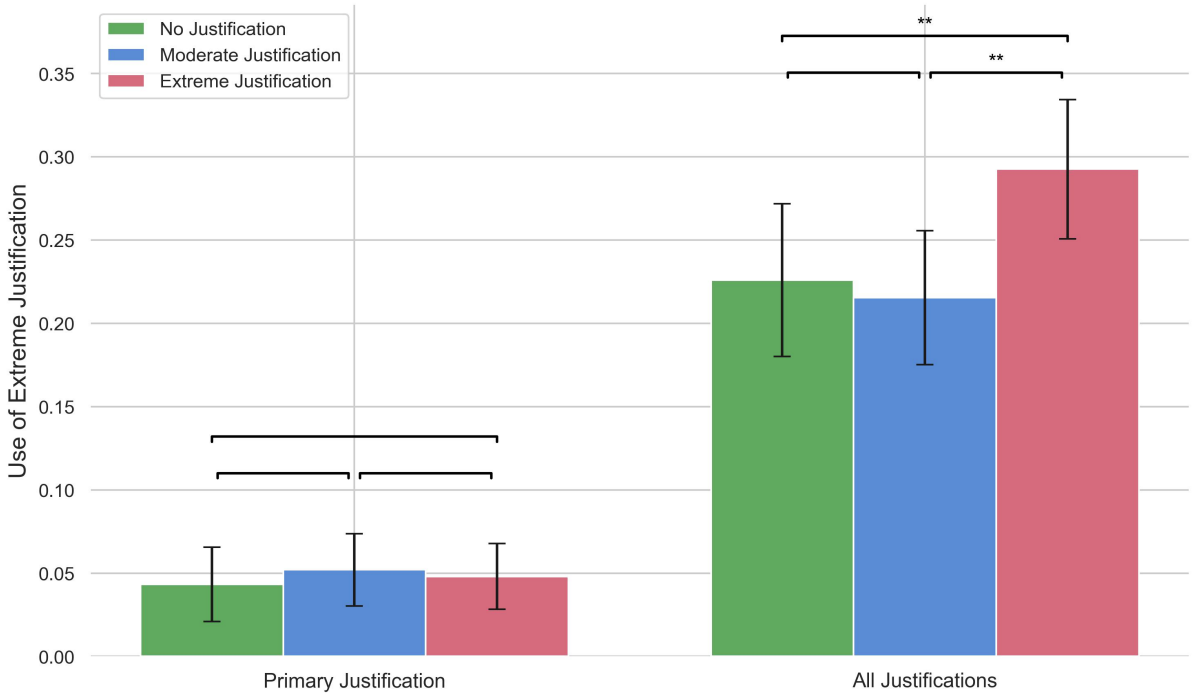
Figure B6: The Effects of Others' Behavior for Different Numbers of Peers



Notes: 'Outgroup Discrimination" is a binary variable equal to 1 if the unequal split was chosen. The figure show "Outgroup Discrimination" as a function of how many of the peers chose to discriminate for the "No Justification" treatment (left graph), "Moderate Justification" (middle graph), and "Extreme Justification" (right graph) treatments, respectively. The dashed yellow line shows the level of discrimination in the "Baseline" treatment. The vertical error ranges represent 95%, confidence intervals.

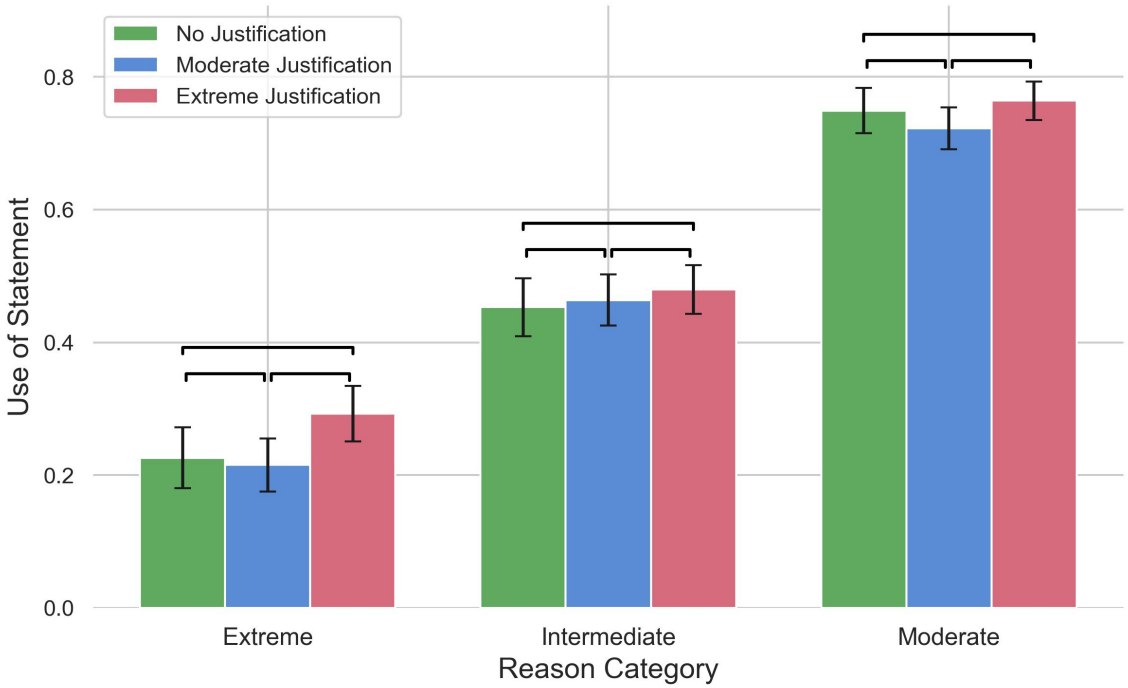
B.5 The Spread of Justifications

Figure B7: The Spread of Extreme Justifications among those who Discriminate



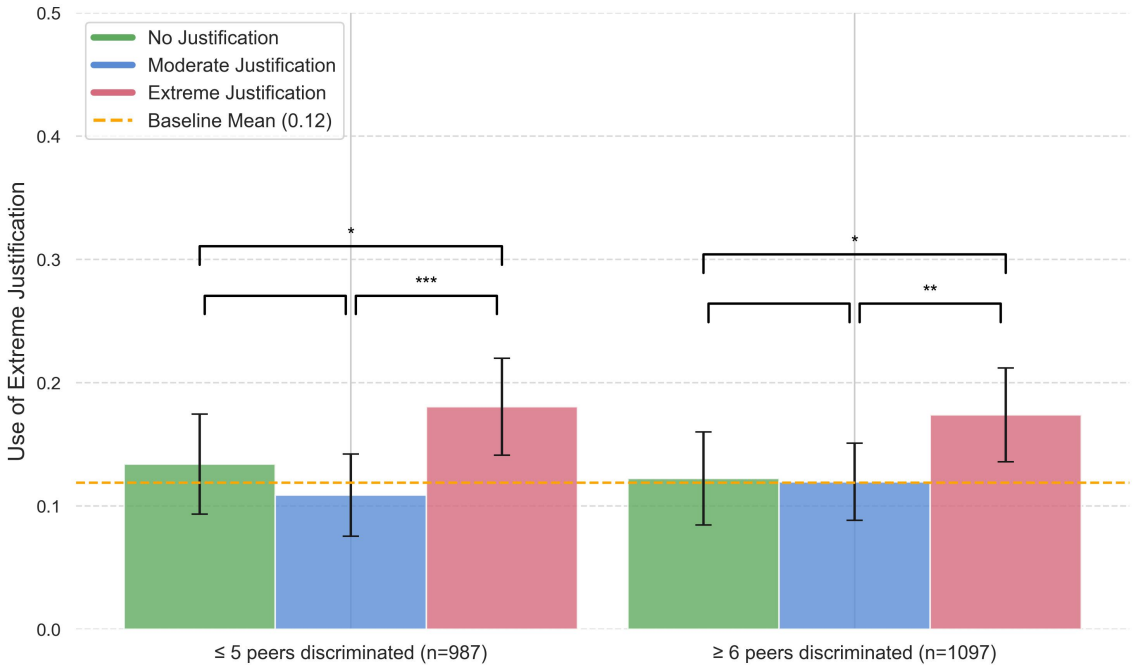
Notes: "Use of Extreme Justification" is a binary indicator equal to 1 if the most extreme statement was used as the primary justification (left three bars) or was selected as one of the justifications with which the subject agrees (right three bars). The graph shows results only for those subjects who discriminated (i.e., results are conditional on discriminating). Significance stars are based on regression specifications as shown in columns (3) and (4) of Table 2. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01

Figure B8: The Spread of Primary Justifications among Discriminators



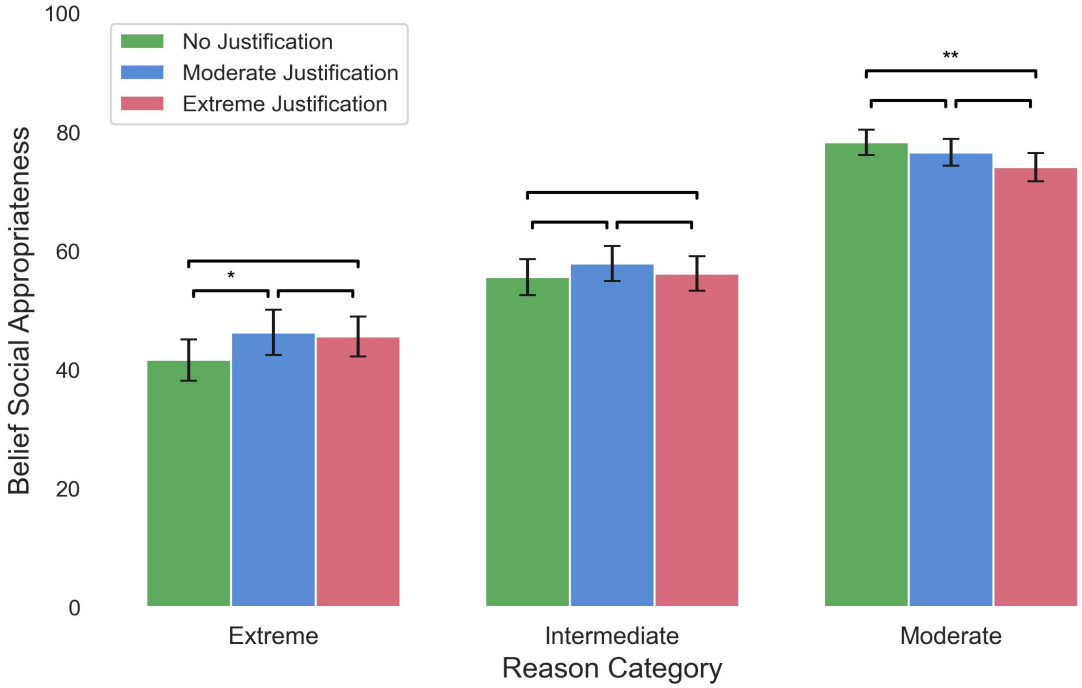
Notes: "Use of Statement" is a binary indicator equal to 1 if a subject used a statement from the associated category as their primary justification for discriminating, split by treatment (for statements in each category, see Appendix Table B1). Significance stars are based on regression specifications following the structure of Table 2. The vertical error ranges represent 95% confidence intervals. *p<0.1, **p<0.05, ***p<0.01.

Figure B9: The Spread of Extreme Justifications across Others' Behavior



Notes: "Use of Extreme Justification" is a binary indicator equal to 1 if the most extreme statement was selected as one of the justifications with which the subject agrees. Results are split by treatment. The three bars on the left include the subset of subjects who were randomly chosen to be shown ≤ 5 peers discriminating and the three bars on the right show results for subjects who were randomly chosen to be shown ≥ 6 peers discriminating. The dashed yellow line shows the average use of the extreme justification in the "Baseline" treatment. Significance stars are based on regression specifications based off of column (2) in Table 2. The vertical error ranges represent 95% confidence intervals. *p< 0.1, **p< 0.05, ***p< 0.01

Figure B10: Belief about Social Appropriateness of Statements



Notes: The figure shows average beliefs (as measured by the midpoint of the slider range placed by subjects) about the social appropriateness of making the different justification statements of each type in each treatment. For the statements in each category, see Appendix Table B1). These beliefs were elicited in an incentivized way (see discussion in Section 2). All data is from wave 2 and results come from regressions that include dummies for the number of peers who discriminate. The vertical error ranges represent 95% confidence intervals. *p<0.1, **p<0.05, ***p<0.01.