**Preregistered Replication and Extension of "Moral Hypocrisy:**

**Social Groups and the Flexibility of Virtue”**

**Claire E. Robertson1, Madison Akles1, & Jay J. Van Bavel1,2\***

**New York University**

**Corresponding author: Jay J. Van Bavel**

1 Department of Psychology, New York University, New York, NY 10003

2 Center for Neural Science, New York University, New York, NY 10003

**Acknowledgements:** We would like to thank the Social Identity and Morality Lab for their helpful comments on this proposal.

TODO:

1. ~~Finish Discussion~~
2. Make Descriptive Stats Tables
3. Fix tense issues in manuscript and highlight.
4. ~~Check equivalence testing~~
5. Complier Average Causal Effects
6. Write up in supplement about Recruitment
7. Fix abstract& Statement of relevance
8. Move pilot info into the supplement
9. ~~Perhaps delete the info about the paradigm itself.~~
10. Make nice graphs for Collective ID
11. Label Supplement at minimum.

Make graphs look nicer if time

# Abstract

The tendency for people to consider themselves morally good while behaving selfishly is known as “moral hypocrisy”. Influential work by Valdesolo & DeSteno (2007) found evidence for intergroup moral hypocrisy, such that people are more forgiving of transgressions when they are committed by an in-group member than an out-group member. We propose a direct and conceptual replication of this work on moral hypocrisy and group membership. We plan to directly replicate their original study in an online paradigm using minimal groups, as well as conceptually replicate and extend their work using natural groups (i.e., political party affiliation). Our experiment will have implications for understanding moral hypocrisy, intergroup bias, and partisanship.

**Keywords**: morality, groups, identity, minimal groups, partisanship

**Statement of Relevance**

Social identities and group memberships exert powerful influences on perception, judgment, and decision-making. Indeed, social identity influences moral decision making, such that people are motivated to forgive moral transgressions perpetrated by their in-group members, while condemning similar transgressions from outgroup members (Valdesolo & DeSteno, 2007). The present research seeks to replicate this finding from Valdesolo & DeSteno (2007), who used artificial minimal groups to test group moral hypocrisy. We will also extend the original research by examining moral hypocrisy among political parties where moral hypocrisy appears to be quite prevalent. This work contributes both to open science replication efforts, and to growing literature on moral hypocrisy in real-world groups.

# Preregistered Replication and Extension of "Moral Hypocrisy:

# Social Groups and the Flexibility of Virtue”

*"Hypocrisy is not a way of getting back to the moral high ground. Pretending you're moral, saying you're moral is not the same as acting morally." -Alan Dershowitz*

# 1.0 Introduction

Although most people hold themselves to a moral code, they are also able to commit immoral acts (Hofmann et al. 2014). These acts may range from the mundane, such as cutting in line, to the extreme, such as shocking someone nearly to death (Milgram, 1963). Regardless, people continue to consider themselves moral beings even after committing immoral acts -- a phenomenon termed Moral Hypocrisy (Batson et al., 1997, Batson et al., 2002). Moral hypocrisy may stem from the psychological need to reframe one’s own immoral actions to justify self identification as moral beings (Shalvi et al., 2011; Shalvi et al., 2015). Critically, however, both one’s moral sense and moral identity are heavily influenced by the social groups they belong to (Graham et al., 2009; Van Bavel et al., in press). Social groups exert a powerful influence, such that people are likely to favor ingroup members and derogate outgroup members (Balliet et al., 2014; Leach et al., 2003; Rathje et al., 2021; Tajfel & Turner, 1979). Therefore, in the current research we ask the question: does moral hypocrisy extend beyond individuals to groups?

Influential work in this area has found that group identity shapes moral hypocrisy (Valdesolo & DeSteno, 2007). Using an elegant study design, researchers found that the same immoral action (assigning an easy task to oneself and an onerous task to someone else) was judged to be more fair when the participant themselves, or a member of the participant’s ingroup, was the perpetrator. It was seen as more unfair when the same action was perpetrated by an out-group member. In other words, the benefits of moral hypocrisy that people allow themselves also extend to in-group members, but not to out-group members.

The implications of these findings are far-reaching; this work suggests that judgements of others' moral character are susceptible to intergroup bias. Moral hypocrisy has strong social consequences, such that those who are viewed as hypocritical deserve more punishment for a transgression compared to non-hypocrites (Effron et al., 2018, Barden et al., 2005). Moral hypocrites are also seen as freeriders because they outwardly signal their own purported morality to gain social status, but do not incur the costs of truly behaving morally (Jordan et al., 2017; Tosi & Warmke, 2020). In the political realm, politicians’ moral hypocrisy reduces judgements of their competency, and elicits negative emotions such as anger from constituents (McDermott et al. 2015; von Sikorski & Herbst, 2020). Thus, real-world moral hypocrisy may heighten negative emotions and contribute to affective political polarization (Finkel et al., 2020).

We seek to replicate Valdesolo and DeSteno’s findings to determine if they generalize to a new sample and a different social identity over a decade later. We will improve the methodology in three main ways: (1) increasing the sample size and statistical power, (2) adding new explanatory analyses, and (3) extending the finding to real world groups to evaluate external validity. First, the original paper has a relatively small sample size: the total sample size in the study was 76, split into four conditions (providing 19 participants in each cell). The reported effect size was *d* = 1.11, which is large for a social psychological study where the average effect size is closer to *d* = .4 (Richard et al., 2003). In the original paper, the researchers also excluded participants who behaved fairly or altruistically. However, subsetting after random assignment based on participant responses is not statistically sound without robustness checks (Lechin, 2000). Thus, we will adopt an intent-to-treat methodology in which all participants are included in our analyses. For robustness, we will also replicate the statistics from the original study using listwise deletion of the altruists, and additionally examine the complier average causal effect (CACE) for where altruistic participants are treated as non-compliers (Hewitt et al., 2006; McCoy, 2017).

Second, the original experiment used a minimal groups procedure–where people are assigned into arbitrary groups (Tajfel et al., 1971). While a minimal groups design offers a well controlled test of the moral hypocrisy effect, it is unclear if intergroup moral hypocrisy would generalize to real world groups, where moral hypocrisy appears to be quite prevalent (Cottle, 2021; Wolsky, 2020). Thus, we will replicate the effect in both minimal groups and natural groups to increase external validity. In everyday life, moral conflict (including moral hypocrisy) is most likely to occur between groups that have historical and/or sociological origins such as religion (Ginges et al., 2007) or political affiliation (Brady et al., 2020; Finkel et al., 2020). Prior research in the lab and in real life suggests that people hold moral double standards regarding their political ingroups and outgroups (Eriksson et al.,2019; Claassen & Ensley, 2016; Solomon et al., 2019). Furthermore, people are more likely to downplay an in-group member’s moral transgressions when they themselves are highly identified with the group (Iyer et al., 2012). Therefore, we will conduct a novel experiment in which partisans will be separated based on their political party identification (i.e., Democrats or Republicans). This will contribute to growing research on moral hypocrisy in real-world groups (McDermott et al., 2015; von Sikorski & Herbst, 2020; Wolsky, 2022).

Third, we will examine the moderating effect of strength of collective identification on intergroup moral hypocrisy. Prior work suggests that the strength of one’s identification with their in-group is associated with increased perceived in-group homogeneity and out-group derogation (Branscombe et al., 1999; Hornsey, 2008; Leach et al., 2003). For example, people judge out-group behaviors more harshly when they are high in collective narcissism – a defensive belief about one’s own in-group’s greatness (Bocian et al., 2021). Therefore, we will examine whether one’s level of collective identification is related to intergroup moral hypocrisy.

We will modify the experiment to occur online to obtain a sufficiently large sample size. Thus, this replication will also explore whether the moral hypocrisy effect will be induced in an online context. While the original study used two confederates to deceive participants into believing that they were interacting with other participants, we will use a real online chat room where participants will interact with three other participants. This change has a number of benefits. First, it will eliminate the need for confederates. Second, we will label each participant with their group identity (i.e. participants are named Overestimator-A or Underestimator-B) to make their group membership salient. Similar measures were taken in the original study, but not included in the original manuscript due to the short report format[[1]](#footnote-1). Otherwise, we propose following the original procedure almost exactly.

**Overview**

In two experiments, we plan to replicate and extend the original research on intergroup moral hypocrisy (Valdesolo & DeSteno, 2007). In the original paper, the authors performed a contrast analysis that examined whether people who made judgements about themselves or their in-groups were significantly more fair than people who judged unaffiliated others or out-group members. This is the main analysis that we will replicate in our experiment. Thus, we hypothesize that (1) people will evaluate themselves and their in-groups more fairly than unaffiliated others and out-group members. In order for the original study to be replicated, H1 must be confirmed in Experiment 1 (with minimal groups). In addition to the original analysis, we hypothesize that (2) people’s evaluations of their own fairness will be greater compared to their evaluations of others’ fairness after committing the same moral transgression. We further hypothesize that (3) people will evaluate their in-group members as more fair than outgroup members after committing the same moral transgressions when the “Self” and “Other” conditions are excluded from analysis.

We also propose some key extensions to the original research. We hypothesize that (4) H1, H2 and H3 will be confirmed when in-groups and out-groups are defined by both minimal groups and natural groups (political party identification). We also hypothesize that (5) the group-based moral hypocrisy effect (H3) will be stronger for natural groups than minimal groups. Finally, we hypothesize that (6) the strength of collective identification will moderate moral hypocrisy, such that people who are strongly identified with their political in-group (6a) will rate their in-group member’s actions as more fair and (6b) their out-group member’s actions as less fair than people who are weakly identified. However, we could find a “black sheep effect” by which people who are highly identified with their groups judge ingroup members more harshly for committing a moral transgression (Marques et al., 1988). This effect may also depend on political ideology – prior work suggests that conservatism is associated with outgroup punishment, whereas liberalism is associated with ingroup punishment (Leshin et al., 2022). Thus, we will examine the effect of both political extremism and political ideology on the level of ingroup and outgroup punishment.

# 2.0 Methods

## 2.1 Participants

To increase the statistical power from the original paper, we plan to increase the sample size in both Experiment 1 and Experiment 2 (see Brandt et al., 2014). To find out how many participants, we conducted a power simulation in R, assuming the average medium effect size common in social psychology, *d* = .4 (Richard et al., 2003; Lovakov & Agadullina, 2021). The original study reported an effect size of 1.11 which would be considered very large for a psychology study (Cohen, 1992). Because we are replicating the experiment in an online context where the manipulation may be less impactful, we chose to be conservative in our effect size estimate.

Because our proposed analyses are multi-stepped and have complex decision rules, we calculated power based on simulated data[[2]](#footnote-2). First we simulated data based on the means from the original study with an effect size one-third the size of the original study (i.e. d = .4 rather than 1.11). We then ran the contrast analysis for H3 on our simulated data and recorded the *p*-value and effect size. We repeated this process 1000 times for various sample sizes. Assuming *d* = .4, with 520 total participants we would achieve 92% power with an alpha of .05 to detect significant differences in our planned contrasts. We also ran a power analysis for equivalence testing (specifically, TOST; two one-sided t-tests) using the TOSTR R package function ‘powerTOSTone.’ We tested whether we would have the power to reject the presence of effects of *d* > 0.2. According to this power analysis, with an alpha of .05 and the proposed sample size of 520, we would have .97 power.

Therefore, each cell in both Experiment 1 and Experiment 2 will have n = 130. We plan to achieve a sample size of 520 using the online survey platform Prolific due to their large survey population, superior researcher controls and data quality. Because of attention check failure, we will collect N = 600 participants in both Experiment 1 and Experiment 2. If, after collecting 600 participants we have not reached 520 participants who passed the attention check and completed the survey, we will continue to recruit in batches of 80 participants until we reach at least 520 participants.

To be eligible for our experiment, participants must be over the age of 18 and reside in the United States. Participants will be paid for 20 minutes of their time at $10 an hour (above federal minimum wage), such that each participant will earn $3.33. Participants will sign up for a timeslot to participate, and will be compensated upon completion of the experiment. We will recruit a nationally representative (on multiple demographic criteria) sample of adults in Experiment 1, and will recruit a politically balanced sample (i.e. 50% Democrats, 50% Republicans) in Experiment 2, reducing the Democratic bias in many online survey groups (Huff & Tingley, 2015). Participants must have above a 90% approval rating on Prolific in order to participate to ensure we have high quality participants.

## 2.2 Experiment 1 Procedure

We will use a procedure that matches the original paper as closely as possible in an online setting. As in the original study, participants will begin the survey by completing the minimal group overestimator/underestimator task with false feedback (Tajfel et al., 1971). The estimation task will consist of participants viewing an array of dots for three seconds, after which they will be prompted to estimate how many dots were in the array. All participants will be shown the same array of dots, but the designation of overestimator or underestimator will be assigned randomly. This will create an in-group and out-group for participants.

In the original study, participants engaged face-to-face with two confederates in a lab. Believing that other people are true actors in the experiment is crucial for psychological induction. Therefore, we will include a chat room where participants interact with three people from their assigned minimal in-group and out-groups. This will be implemented into the qualtrics survey itself using [Smarttriqs](https://smartriqs.com/) (Molnar, 2019). In this phase, the participants’ will be assigned chat names that correspond to the minimal group to which they are assigned (ex. “Overestimator-1”). The roles are Overestimator-1, Underestimator-2, Overestimator-3, and Underestimator-4. Participants group assignments will be labeled in order to further strengthen the minimal groups induction.

Participants will be entered into a chatroom with other participants, with the prompt “Please take the next few minutes to chat with other participants about being an overestimator or underestimator. Remember, you may have seen different images.” They will not be able to progress in the survey until 60 seconds have passed. They will be told they may have seen different images because participants were assigned to be Overestimators or Underestimators randomly. Thus, some participants who estimate higher numbers of dots may be told they are Underestimators, while someone who guesses a lower number of dots may be told they are an Overestimator.

After participating in the chat, participants will be asked to report their collective identification with their in-group and compared to their out-group (Van Bavel & Cunningham, 2012). Participants will be asked to respond to the following three items for each group: “I value being a member of the [Overestimator/Underestimator] group.” “I am proud to be a member of the [Overestimator/Underestimator] group,” and “Being a member of the [Overestimator/Underestimator] group is an important part of my identity.” (Van Bavel & Cunningham, 2012). Collective identification will be calculated by taking the difference of participants' in-group scores and out-group scores, such that positive scores reflect greater collective identification.

Participants in all conditions will then read instructions stating that researchers are interested in performance on two different tasks**.** Task 1 (the “green” task in the original paper) is a simple task consisting of a photo hunt game where participants will be prompted to ‘spot the difference’ between two images and a short personality inventory, and is designed to be fun. Task 2 (the “red” task in the original paper) is a complex task consisting of mental rotation on an irregular block shape and logic problems from the LSAT, and is designed to be cognitively taxing. In order to make the differences between the tasks salient, participants will be given examples of both the green task and the red task to complete, along with feedback on their accuracy. In line with the original study, we will also tell participants that the green task will take approximately 8 minutes, while the red task will take approximately 20 minutes.

In the original study, the participants were in a lab and participating for course credit, where the length of the study had no effect on the possible compensation for their participation. However, on most online survey platforms participants are compensated per minute for their task participation. This might have added an incentive for people to choose the red task in order to earn more money, which could interfere with whether choosing the green task for one’s self was seen as immoral. In order to mitigate this potential confound, we will tell participants that they will be paid for 20 minutes of work regardless of which task they participate in. Thus, the green task is still a more desirable task to participate in.

Participants will also be given an attention check in this phase of the experiment, before experimental random assignment. The attention check is designed to look like a regular survey question, and reads: “There are lots of different types of questions that we may ask participants. Some types of questions look at personality, whereas others may test certain sets of skills. Others may test to ensure that participants read the entire question. Please select ‘Somewhat Disagree’ from the selection below.” Participants must select “Somewhat Disagree” on a 1 (Strongly Disagree) to 7 (Strongly Agree) Likert scale to pass the attention check. Because this attention check will occur before participants are assigned to experimental conditions, it will not be impacted by conditional drop out.

After completing the sample tasks, participants will be told that some participants will be selected to participate in the green task, while others will participate in the red task. Participants will be told that, in order to keep the researchers blind to the condition of each participant, the researchers are using a newly developed assignment procedure in which a random subset of participants will be allowed to choose which task they want to complete. Whatever task they do not complete will be assigned to another participant. Those chosen to make assignments can either assign tasks randomly by using a computer randomizer, or they can select one task for themselves, leaving a future participant to complete the other, unselected task. Therefore, a participant who chooses to assign themselves to the faster, easier green task is forcing a different participant to complete the longer, harder red task. This is how we, and the original paper, operationalize a moral transgression.

At the phase of the experiment where task selection occurs, the participants will be split into four possible conditions, matching the original paper. In the “Self” Condition, participants will be instructed to select which task they would like to complete. In the original paper, 17 out of 19 participants in the "Self" Condition assigned themselves the less onerous task. The two participants who chose to act altruistically (one using the computer randomizer and one choosing the worse task for themselves) were excluded from analyses. Scholars caution against subsetting or excluding participants based on task choices (Lachin, 2000). Thus, we will use an intent-to-treat methodology in which all participants will be analyzed according to the condition they were randomized into (McCoy, 2017).

In the three remaining conditions – the “Other” Condition, the “In-group” Condition, and the “Out-group” Condition – participants will be told that another participant has been selected to assign the tasks. In the “Other” Condition, no other information will be given about this other participant, matching the “unaffiliated other” condition from the original paper. In the “In-group” Condition, participants will be told that the participant they observe assigning tasks is part of the same minimal group that the participant is. Participants will read the following information: “[Overestimator-2/Underestimator-4] has been randomly selected to assign roles.” Participants are then reminded of the two tasks, and are then told to wait while the [Overestimator/Underestimator] are making their choices. Participants will see the label that is the same as the label they were assigned during the minimal groups task, matching the “ingroup other” condition from the original paper. In the “Out-group” Condition, participants will be told that the observed participant is a member of their minimal out-group, matching the “outgroup other” condition from the original paper.

Participants in these three conditions will then be told that they will learn of that participant’s decision. Participants will be reminded that the allocator has the choice to use a randomizer, or to assign themselves to one of the tasks. After a brief waiting period, participants will learn via experimenter generated false feedback that the other participant chose to behave selfishly by assigning themselves the “green” task and assigning a future participant the “red” task.

Once task selection is completed, participants will answer questions about the “experimenter blind” selection procedure. Embedded in the questionnaire is our question of interest: “How fairly did the other participant act when assigning the tasks?” This is the dependent measure of interest, and will be answered on a 7-point Likert scale (from 1 = “*extremely unfairly*” to 7 = “*extremely fairly*”). They are also asked three distractor questions: “How likely do you think people are to assign themselves the green task?” on a 10-point Likert scale (from 1 = “*extremely unlikely*” to 7 = “*extremely likely,*” “Do you think the assignment procedure for tasks is blind (e.g. the researchers are not assigning tasks)” and “Do you have any other feedback on the new "experimenter-blind" task assignment procedure?”).

## 2.3 Experiment 2 Procedure

The procedure for Experiment 2 is identical to Experiment 1, except that instead of using minimal groups to establish participants’ in-group and out-group membership, we will use participants' pre-existing political party identities. Political party identities will be taken from people’s Prolific battery. Participants' chat room names will be reflective of their political party identification (i.e. Democrat-1, Republican-2, Democrat-3, and Republican-4). Participants will still converse about the minimal groups paradigm in the chatroom.

Participants in the "Self" Condition will follow the exact same procedure for the "Self" Condition laid out in Experiment 1. The “Other” Condition will follow exactly the same procedure as in Experiment 1, where they learn that another participant behaved selfishly and learn nothing about that participant’s identity. In the “In-group” Condition, participants will learn that someone from their political in-group behaved selfishly (e.g., a participant who identifies as a Republican will be told they are learning of a Republican’s decision). In the “Out-group” Condition, participants will learn that someone from their political out-group behaved selfishly (e.g., a participant who identifies as a Republican will be told they are learning of a Democrat’s decision). To inform participants of the target’s political in-group status, the participant will read the instructions as follows: “[Player 1 (Democrat)/Player 2 (Republican)/Player 3 (Democrat)/Player 4 (Republican)] has been randomly assigned to assign roles.” Participants will then be asked the same question regarding how fair they thought the other participant acted, embedded in a series of distractor questions.

# 3.0 Analysis Plan

For the main analyses, participants who fail the attention check or who do not complete the experiment will be excluded. The attention check will appear before random assignment to condition to avoid post-treatment bias (Montgomery et al., 2018). In the procedure in the original research, participants in the "Self" Condition who make altruistic choices (i.e. choose the randomizer or the red task for themselves) were excluded. However, because conditioning inclusion on a post-treatment variable can violate random assignment due to nonrandom attrition (Montgomery et al., 2018), we will conduct intent-to-treat analyses (ITT) for all of our main measures in which all "Self" Condition participants will be included, regardless of whether they made an altruistic choice. For robustness, we will conduct two additional analyses in the supplemental information. First, we will estimate the complier average causal effect (CACE) for those in the “Self” Condition (Montgomery et al., 2018). This procedure attempts to estimate the effect only among those who received the treatment from the experiment. Second, we will run an analysis in which we exclude all altruistic participants from analyses via listwise deletion to match the original paper. For results to successfully replicate the original study, the hypotheses must be supported by the intent-to-treat analysis. For all of the proposed analyses, we will be using alpha = 0.05 for significance testing. Regarding concerns about floor or ceiling effects in statistical analyses, there is no reason in the original paper to suspect that floor or ceiling effects will occur, nor do we find evidence of floor or ceiling effects in Pilot Experiment 1 or Pilot Experiment 2 (see Supplemental Information S1.0).

## 3.1 Experiment 1 Analysis Plan

In Experiment 1, we will test H1, H2, H3 and H6. All proposed analysis code can be found [here](https://github.com/clairerobertson/moral_hypocrisy_replication). Hypothesis 1 predicts that people will rate themselves and their in-group members as behaving more fairly than their out-group members or unaffiliated others. To test this, we will use a planned contrast where the "Self" Condition and the “In-group” Condition have contrast weights of 1 and the “Other” Condition and the “Out-group” Condition have contrast weights of -1, matching the original study. If the contrast analysis is significant and the mean fairness ratings for the self and in-group members are greater than the mean fairness ratings for out-group members and unaffiliated others, then we can conclude that H1 is supported.

If this contrast analysis does not reach the alpha = 0.05, we will conduct equivalence testing to examine whether we have an absence of a meaningful effect. We will use the TOSTR package from *R*, with the function powerTOSTone. If the equivalence test is significant, we can conclude there is no meaningful effect in our data.

Hypothesis 2 predicts that people will rate themselves as behaving more fairly than others. To test this, we will use a planned contrast where the "Self" Condition has a contrast weight of 3 and the “Other”, “In-group”, and “Out-group” Conditions have contrast weights of -1. If the contrast analysis is significant and the fairness rating for the self is higher than the mean fairness ratings in the other conditions, then we can conclude that H2 is supported.

Hypothesis 3 predicts that people will rate in-group members as behaving more fairly than out-group members. To test this, we will use a planned contrast where the “In-group” Condition has a contrast weight of 1, the “Out-group” Condition has a contrast weight of -1, and the “Self” and “Other” Conditions have contrast weights of 0. If the contrast analysis is significant and the fairness rating for in-group members is higher than the fairness ratings for out-group members, then we can conclude that H3 is supported.

Hypothesis 6 predicts that the strength of collective identification will moderate moral hypocrisy, such that people who are strongly identified with their in-group (6a) will rate their in-group member’s actions as more fair and (6b) their out-group member’s actions as less fair. To test this, we will use a multiple linear regression model, in which we regress fairness ratings on dummy-coded condition (the “In-group” Condition and the “Out-group” Condition), mean collective identification, and their interaction. If the interaction term is significant such that fairness ratings for in-group members increase as collective identification increases, while fairness ratings decrease for out-group members as collective identification increases, then we can conclude that H6 is supported

## 3.2 Experiment 2 Analysis Plan

In Experiment 2, we will test H4, H5 and H6. Hypothesis 4 states that Hypotheses 1, 2, and 3 will be confirmed in natural groups as well as miniman groups. Thus, we will repeat all analyses proposed for H1, H2 and H3 on the natural groups sample. If Hypothesis 1, 2, and 3 are supported when groups are based on political party and minimal group assignment, then H4 is supported.

Hypothesis 5 states that the effects of moral hypocrisy will be larger for natural groups compared to minimal groups. To test this, we will conduct a 2 (contrast: “In-group” vs. “Out-group”) X 2 (Experiment: Minimal Groups vs. Natural Groups) ANOVA. We expect that the interaction term will be significant, and that simple effects will reveal that in-group fairness will be higher for natural groups compared to minimal groups, and outg-roup fairness will be lower for natural groups compared to minimal groups. If the above predictions are all supported, then we can conclude that H5 is supported.

Hypothesis 6 states that the strength of collective identification will moderate moral hypocrisy, such that people who are strongly identified with their political in-group (6a) will rate their in-group member’s actions as more fair and (6b) their out-group member’s actions as less fair. To test this, we will use a multiple linear regression model, in which we regress fairness ratings on dummy-coded condition (“In-group” Condition and “Out-group” Condition), mean collective identification, and their interaction. If the interaction term is significant such that fairness ratings for in-group members increase as collective identification increases, we will investigate the nature of the relationship using a follow-up simple slopes analysis. We will conduct *post hoc* tests of the relationship between fairness ratings and collective identification for those in the “In-group” Condition and the “Out-group” Condition) separately. If the slope is positive for those judging their in-group members moral transgressions, and negative (or flat) for those judging their out-group members moral transgressions, then we can conclude that H6 is supported.

We will also conduct two robustness tests of our manipulations. First, we will examine whether those who were assigned to be overestimators and those who were assigned to be underestimators significantly differ in judgements across our four conditions. We hypothesize that there will be little to no difference between overestimators and underestimators judgements in the same conditions. To test this, we will conduct equivalence testing across the four conditions comparing responses from overestimators to underestimators in the same conditions. Using established guidelines from Cohen (1992) and procedures from Lakens (2017) we will consider *d* = 0.2 our smallest effect size of interest. We hypothesize that the effect of being in the over/underestimator group will not be statistically different from zero in any of the four conditions. We will run this robustness check for both Experiment 1 and Experiment 2.

Second, we will also analyze the levels of identification with the in-group and out-group to ensure the minimal group manipulation is inducing people to identify more with their minimal in-group. In previous work, we found that there was a clear difference in identification between these minimal groups, with people identifying more with their in-group compared to the out-group (*d* = .91; Van Bavel & Cunningham, 2012). We will use a one-samples t-test on participants’ collective identification difference scores to examine whether participants’ difference scores are significantly different from chance. We hypothesize that participants' difference scores will be significantly higher than zero, indicating that they feel more identified with their in-group compared to their out-group. We expect this to be true for both minimal groups and natural groups.

**Results:**

**Study 1:**

We recruited 610 American, nationally representative participants from Prolific[[3]](#footnote-3). After removing participants who failed attention checks (*n* = 6), who accidentally took the survey more than once (*n* = 9), and those who did not consent for us to use our data (*n* = 5), our final sample consisted of 590 participants (*Mage* = 35.68 years *SDage* = 14.61, 295 men, 275 women). Full gender and ethnicity breakdown of the sample can be found in the Supplement XXXX.

We preregistered an intent-to-treat analysis, in which we include both altruists and moral transgressors in the Condition 1 sample. However, the altruists in our sample rated themselves as behaving significantly more fairly (*Mfairness* = 6.41) than the transgressors in our sample, *Mfairness* = 4.08, *t*(146) = 11.58, *p* < 0.001. This introduced a major confound, as the altruists are judging the fairness of a fair decision, while all other participants are judging the fairness of an unfair decision, which made results including altruists difficult to interpret. Thus, we made a post hoc decision to report results excluding the altruists in the main text (Figure 1A) along with the intent-to-treat analyses (Figure 1B). This decision was also made by the original authors, and thus closely replicates as in the original study. We excluded participants in Condition 1 who chose to behave fairly (i.e. used the randomizer, *n* = 47) or altruistically (chose the red task for themselves, *n* = 4).

A diagram of different shapes

Description automatically generated

Figure 1:

We aimed to replicate the findings from the original paper. First, we predicted that people will evaluate themselves and their in-groups more fairly than unaffiliated others and out-group members. For minimal groups, this was not significant, *F*(1, 536) = 0.152, *p* = 0.696. Results did not change when altruists were included in the sample *F*(1, 592) = 0.198, *p* = 0.656. We conducted equivalence testing to determine if the nonsignificant contrast analysis was smaller than the smallest effect size of interest. We used the R package TOSTER and the function equ\_ftest. For the “Self & In-group” vs. “Other & Out-group” contrast, where *F*(1, 536) = 0.152, the 95% CI of our effect size was [0.000, 0.0096] which does not contain our a priori equivalence bounds of -0.2 and 0.2. On the basis of the null hypothesis test and the equivalence testing combined, we can conclude that the observed contrast effect is statistically not different from zero, and is statistically equivalent to zero.

Second, we predicted that people’s evaluations of their own fairness will be greater compared to their evaluations of others’ fairness after committing the same moral transgression. Again, for natural groups, we found that this comparison was not significant when altruists were excluded, *F*(1, 536) = 1.2661, *p* = 0.261 When altruists were included, we found that those in the “Self” condition rated themselves as having behaved significantly more fairly than those in conditions judging others *F*(1, 592) = 40.165, *p* < 0.001. Third, we hypothesized that people would evaluate their in-group members as behaving more fairly than outgroup members after committing the same moral transgressions when the “Self” and “Other” conditions are excluded from analysis. For natural groups, we found that people did not judge ingroup and outgroup members significantly differently *F*(1, 536) = 3.219, *p* = 0.073. Results did not change when altruists were included in the sample *F*(1, 592) = 2.883, *p* =0.090.

Next, we examined whether the strength of collective identification moderated moral hypocrisy. We predicted that people who were strongly identified with their groups would rate their ingroup members actions as more fair and their outgroup members actions as less fair. We found a significant main effect for collective identification 𝛃 = 0.298, t(290) = 3.79, p < 0.001, and a significant interaction effect between condition and collective identification 𝛃 = -0.393, *t*(290) = -3.64, *p* < 0.001. Simple slopes analysis revealed a significant effect in the ingroup condition *t*(290)= 3.79, *p* < 0.001, such that participants judged other ingroup members more fairly the more they were identified with their minimal group memberships.

We conducted equivalence testing to determine if the nonsignificant contrast analyses were smaller than the smallest effect size of interest. We used the R package TOSTER and the function equ\_ftest. For our main contrast of interest

**Study 2:**

We recruited 606 American participants from Prolific, 50% Democrats and 50% Republicans. After removing participants who failed attention checks (*n* = 8), who accidentally took the survey more than once (*n* = 13), and those who did not consent for us to use our data (*n* = 8), our final sample consisted of 577 participants (*Mage* = 33.63 years *SDage* = 13.52, 283 men, 280 women). Full gender and ethnicity breakdown of the sample can be found in the Supplement XXXX. Participants were assigned roles as either Democrats or Republicans based on their reported political party identification on Prolific. In our sample, 50.78% of participants reported being Democrats, and 49.22% reported being Republicans on Prolific. We also asked participants to report their political orientation on a 7-point Likert scale (1 = Very Liberal, 7 = Very Conservative). We found that 281 participants identified as liberal, 260 identified as conservative, and 44 identified as moderate.

Like Study 1, some participants in Condition 1 who chose to behave fairly (i.e. used the randomizer, N = 40) or altruistically (chose the red task for themselves, N = 6), and we found that the altruists in our sample rated themselves as behaving significantly more fairly (*Mfairness* = 6.41) than the transgressors in our sample, *Mfairness* = 4.08, *t*(146) = 11.58, *p* < 0.001. Thus, we make the same post-hoc decision to report results excluding the altruists (Figure 3A) in the main text along with the intent-to-treat analyses (Figure 3B).

A diagram of different shapes and colors

Description automatically generated with medium confidence

Figure 3:

First, we predicted that people would evaluate themselves and their in-groups more fairly than unaffiliated others and out-group members (H1). For natural groups, this was not significant, F(1, 528) = 1.864, p = 0.17. Results did not change when altruists were included in the sample *F*(1, 581) = 1.585, p = 0.21. Second, we predicted that people’s evaluations of their own fairness will be greater compared to their evaluations of others’ fairness after committing the same moral transgression (H2). Again, for natural groups, we found that this comparison was not significant when altruists were excluded, *F*(1, 528) = 1.441, *p* = 0.23. When altruists were included, we found that those in the “Self” condition rated themselves as having behaved significantly more fairly than those in conditions judging others *F*(1, 581) = 40.606, *p* < 0.001. Third, we hypothesized that people would evaluate their in-group members as behaving more fairly than outgroup members after committing the same moral transgressions when the “Self” and “Other” conditions are excluded from analysis (H3). For natural groups, we found that people believed that their ingroup members acted significantly more fairly than their outgroup members *F*(1, 528) = 9.043, *p* = 0.003. This was also true when altruists were included in the sample *F*(1,581) = 7.461, *p* = 0.007.

Next, we examined whether the strength of collective identification moderated moral hypocrisy (H6). We predicted that people who were strongly identified with their political groups would rate their ingroup members actions as more fair and their outgroup members actions as less fair. We did not find significant main effects for either condition or collective identification, or interaction effects interaction effect for these predictions, 𝛃 = -0.019, t(284) = -0.268, p = 0.789.

Finally, we examined whether the group-based moral hypocrisy effect was stronger for natural groups as compared to minimal groups. We did not find significant effect of minimal vs. natural group, *F*(1, 581) = 1.84, *p* = 0.175, nor did we find a significant interaction effect, *F*(1, 581) = 0.59, *p* = 0.443. We did see a significant effect of condition, such that participants judged ingroup members actions as more fair compared to outgroup members, *F*(1, 581) = 11.69, *p* < 0.001.

We also ran an exploratory analysis, which revealed that people reported significantly greater collective identification with their political groups compared to minimal groups, *t*(967.4) = -15.81, *p* < 0.001.

**Discussion:**

Valdesolo and DeSteno (2007) found that people judged themselves and their ingroups as behaving more fairly than unaffiliated others and outgroup members for the same moral transgression (unfairly assigning themselves an easier experimental task), demonstrating that the moral hypocrisy effect extends to ingroups and outgroups. In two studies, we were unable to replicate their findings in both minimal groups and natural groups. However, our patterns of results were consistent with two established psychological phenomena: minimal groups ingroup favoritism CITE), and affective polarization (CITE).

In Study 1, we found that the level of identification with one’s minimal group predicted fairness ratings of ingroup members, but not outgroup members. People who were more identified with their minimal group rated ingroup members immoral behavior as more fair. This is consistent with previous work showing that the minimal groups effect is typically driven by ingroup favoritism, rather than outgroup animosity (CITE).

We did find evidence of outgroup animosity among natural groups, such that people judged political outgroup members as having behaved more unfairly than political ingroup members (CITE). Furthermore, exploratory analyses show that this outgroup animosity effect was present in both Democrats and Republicans. These findings are consistent with affective polarization, (IYENGAR ET AL., 2019), or the dislike and distrust of those from opposing political parties. Importantly, people were demonstrating outgroup animosity towards outpartisans based on non-political opinions, demonstrating how affective polarization can bias perceptions of outpartians general character, even separate from politics (MOORE-BERG ET AL., CIKARA & LEES, ET AL).

For natural groups, we did not find an effect of collective identification on fairness ratings. However, exploratory analyses revealed that people reported significantly greater collective identification with their political groups compared to minimal groups. This reveals a potential explanation for the lack of effects in Study 1 – the minimal groups induction may have only been effective for some participants. Exploratory analyses including only people who were highly identified with their minimal group found an outgroup animosity effect similar to Study 2 with natural groups (See Supplementary Analyses XXXX). This may explain why we find an overall effect of outgroup animosity in natural groups that is not moderated by collective identification – collective identification stronger at baseline in natural groups compared to minimal groups.

We also examined whether participants engaged in individual moral hypocrisy, such that people rate themselves as more moral than others for the same transgression (CITE). We found mixed results. When we included people who made fair or altruistic decisions, we did find that people rated themselves as behaving more fairly than others. However, these results are difficult to interpret, given that participants who chose to use a randomizer to assign tasks, or chose the difficult task for themselves, objectively did behave more fairly than transgressors, who assigned themselves the easier task. When only transgressors were included in our sample, we did not find that participants rated themselves as more moral than others.

Although we failed to replicate the original findings of Valdesolo & DeSteno (2007), the current work provides new evidence of outgroup animosity and affective polarization among political groups, even in non-political contexts. This work further suggests that collective identification is an important predictor of ingroup favoritism in minimal group inductions. Thus, future research may want to focus on differences in ingroup versus outgroup judgements of behavior rather than judgements of the self versus others.

# 

# References

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological bulletin, 140*(6), 1556.

Barden, J., Rucker, D. D., & Petty, R. E. (2005). "Saying one thing and doing another": Examining the impact of event order on hypocrisy judgments of others. *Personality & social psychology bulletin, 31*(11), 1463–1474.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of personality and social psychology*, *72*(6), 1335.

Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330.

Bocian, K., Cichocka, A., & Wojciszke, B. (2021). Moral tribalism: Moral judgments of actions supporting ingroup interests depend on collective narcissism. *Journal of Experimental Social Psychology,* 93, 104098.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science, 15*(4), 978-1010.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. *Social Identity: Context, Commitment, Content,* 35-58

Claassen, R. L., & Ensley, M. J. (2016). Motivated reasoning and yard-sign-stealing partisans: Mine is a likable rogue, yours is a degenerate criminal. *Political Behavior, 38*(2), 317-335.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological bulletin.*

Cottle, M. (2021, May 4). *Who Cares About Hypocrisy?* The New York Times. https://www.nytimes.com/2021/05/04/opinion/republicans-biden-hypocrisy.html

Effron, D., Markus, H., Jackman, L., Muramoto, Y., Muluk, H. (2018). Hypocrisy and culture: Failing to practice what you preach receives harsher interpersonal reactions in independent (vs. interdependent) cultures. *Journal of Experimental Social Psychology, 76,* 371-384.

Eriksson, K., Simpson, B., & Strimling, P. (2019). Political double standards in reliance on moral foundations. *Judgment & Decision Making, 14*(4).

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Druckman, J. N. (2020). Political sectarianism in America. *Science, 370*(6516), 533-536.

Ginges, J., Atran, S., & Medin, D. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Sciences, 104*, 7357-7360.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology, 96*(5), 1029.

Hewitt, C. E., Torgerson, D. J., & Miles, J. N. (2006). Is there another way to take account of noncompliance in randomized controlled trials?. *Cmaj, 175*(4), 347-347.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340-1343.

Hornsey, M. J. (2008). Social identity theory and self-categorization theory: A historical review. *Social and personality psychology compass, 2*(1), 204-222.

Huff, C., & Tingley, D. (2015). “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics, 2*(3), 2053168015604648.

Iyer, A., Jetten, J., & Haslam, S. A. (2012). Sugaring o'er the devil: Moral superiority and group identification help individuals downplay the implications of ingroup rule breaking. *European Journal of Social Psychology, 42*(2), 141-149.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science, 28*(3), 356-368.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials, 21*(3), 167-189.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362.

Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of Personality and Social Psychology, 84*(5), 932.

Leshin, R. A., Yudkin, D. A., Van Bavel, J. J., Kunkel, L., & Rhodes, M. (2022). Parents’ political ideology predicts how their children punish. *Psychological science, 33*(11), 1894-1908.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology, 51*(3), 485-504.

Marques, J. M., & Paez, D. (1994). The ‘black sheep effect’: Social categorization, rejection of ingroup deviates, and perception of group variability. *European review of social psychology, 5*(1), 37-68.

McCoy, C. E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *Western Journal of Emergency Medicine, 18*(6), 1075.

McDermott, M. L., Schwartz, D., & Vallejo, S. (2015). Talking the talk but not walking the walk: Public reactions to hypocrisy in political scandal. *American Politics Research, 43*(6), 952-974.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 6*7(4), 371.

Molnar, A. (2019). SMARTRIQS: A simple method allowing real-time respondent interaction in Qualtrics surveys. *Journal of Behavioral and Experimental Finance, 22*, 161-169.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760-775.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26).

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of general psychology, 7*(4), 331-363.

Shalvi, S., Dana, J., Handgraaf, M., & De Dreu, C.K.W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes, 115*(2), 181-190.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-Serving Justifications: Doing Wrong and Feeling Moral. *Current Directions in Psychological Science, 24*(2), 125–130.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*(4), 267-281.

Solomon, E. D., Hackathorn, J. M., & Crittendon, D. (2019). Judging scandal: Standards or bias in politics. *The Journal of Social Psychology, 159*(1), 61-74.

von Sikorski, C., & Herbst, C. (2020). Not practicing what they preached! Exploring negative spillover effects of news about ex-politicians’ hypocrisy on party attitudes, voting intentions, and political trust. *Media Psychology, 23*(3), 436-460.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology, 1*(2), 149-178.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A reader, 56*(65), 9780203505984-16.

Tosi, J., Warmke, B. (2020). Grandstanding: The Use and Abuse of Moral Talk. *United Kingdom: Oxford University Press.*

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689-690.

Van Bavel, J. J., & Cunningham, W. A. (2012). A social identity approach to person memory: Group membership, collective identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin, 38*(12), 1566–1578.

Van Bavel, J. J., Packer, D. J., Haas, I. J., & Cunningham, W. A. (2012). The importance of moral construal: Moral versus non-moral construal elicits faster, more extreme, universal evaluations of the same actions. *PLOS ONE, 7*(11), e48693.

Van Bavel, J. J., Packer, D., Ray, J. L., Robertson, C., & Ungson, N. D. A. (In Press). How social identity tunes moral cognition. In N. Elemers, S. Pagliaro, & F. Nunspeet (Eds.) *Routledge Handbook of the Psychology of Morality.*

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences, 22*(3), 213-224.

von Sikorski, C., & Herbst, C. (2020). Not practicing what they preached! Exploring negative spillover effects of news about ex-politicians’ hypocrisy on party attitudes, voting intentions, and political trust. *Media Psychology, 23*(3), 436-460.

Wolsky, A. D. (2022). Scandal, Hypocrisy, and Resignation: How Partisanship Shapes Evaluations of Politicians’ Transgressions. *Journal of Experimental Political Science, 9*(1), 74-87.

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychological Inquiry, 27*(4), 255-274.

# 

# Figures

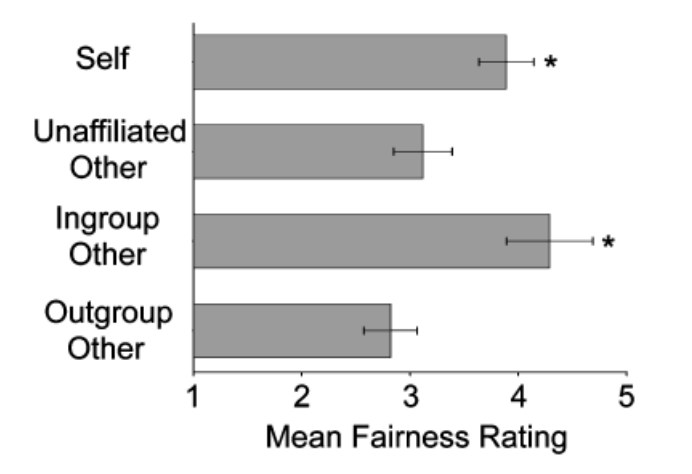
****

Figure 1: Original image from Valdesolo & DeSteno (2007) depicting mean fairness ratings for the self (Condition 1), an unaffiliated other (Condition 2), an ingroup other (Condition 3) and an outgroup other (Condition 4).

# Supplemental Information

## S1.0 Pilot Samples

Due to the sample size requirements that our power analysis revealed, we chose to replicate the original study in an online context. However, we wanted to ensure that our online procedure was true to the original study. Namely, we wanted to ensure that the two critical pieces of the manipulation were effective in an online context. In our two initial pilot experiments (Pilot Experiment 1 and Pilot Experiment 2) using NYU student samples, we administered false feedback via a fake screen recording to show participants the moral transgression of other participants. To make this compelling, first, we needed participants to believe that they were interacting with other, real participants. Secondly, we needed participants to believe that they watched another real participant make a moral transgression. We ran two pilot experiments to test our procedure. The research complies with all relevant ethical regulations. Ethics approval (IRB-FY2022-5934) was obtained from the institutional review board at New York University. Participants in the pilot experiments were recruited from the subject pool of the Department of Psychology at New York University in exchange for 0.5 hrs of research credit for varying psychology courses. In order to ensure that our procedure was viable on our population of interest, we ran two more pilot experiments (Pilot Experiment 3 and Pilot Experiment 4) on the survey platform Prolific. Participants were paid for 20 minutes of their time at $10 an hour (above federal minimum wage), such that each participant earned $3.33. Participants signed up for a timeslot to participate, and were compensated upon completion of the experiment. We recruited a standard U.S. sample in Pilot Experiment 3, and a politically balanced U.S. sample (i.e. 50% Democrats, 50% Republicans) in Pilot Experiment 4.

### *S1.1.0 Pilot Experiment 1*

In Pilot Experiment 1, we collected data from 61 participants via SONA (Department of NYU Psychology Research Participant System). Participants signed up for a prespecified time slot, and were reminded the day of and the day before of their upcoming timeslot. Four participants had to enter the experiment at the same time for the experiment to function.

#### S1.1.1 Pilot 1 Results

In Pilot 1, we were most concerned with validating our experimental procedure. First, we examined whether participants believed they were speaking to real people during the chat portion of the experiment. We found that 58/61 (95%) participants believed they were actually speaking to other participants during the experiment. We then wanted to see whether participants believed they were watching a real person’s screen recording. We found that 33/61 (54%) participants believed they were watching a real person’s screen recording. This ratio was lower than we wanted, so we looked at the narrative reports from participants as to why they did not believe the screen recording. The first recurring feedback we got was that the screen recording was “ready” for their viewing too quickly for it to have come from another participant, and that participants. Other participants reported that they were suspicious that the screen recording was fake because they were never asked to consent to screen recording themselves. Using their feedback, we tweaked our protocol and ran another pilot.

### *S1.2.0 Pilot Experiment 2*

In Pilot Experiment 2 we collected data from 201 participants via SONA (Department of NYU Psychology Research Participant System). Participants signed up for a prespecified time slot, and were reminded the day of and the day before of their upcoming timeslot.

In Pilot 1, we received feedback that participants did not believe that they were watching a screen recording of another real person’s actions. We looked through the written feedback and made several changes to make the screen recording more believable. First, we addressed the concern that the screen recording process did not take enough time to be realistic. In Pilot 2, participants were shown a faux loading screen with the instruction “Screen recording from another participant is currently being uploaded and converted to video for you to view. This may take 30-45 seconds for processing, thank you for your patience” for 45 seconds. Second, we addressed the concern that participants were not asked for their consent to screen record at the beginning of the experiment. Thus, we added a section after the consent form that read “During this study, we may ask permission to ‘screen-record’ your mouse movements. We will prompt you before we start screen recording, and only survey activity will be recorded. Nothing on your desktop or outside of the browser window will be recorded,” and had participants check a box to confirm they understood this. We also added an attention check in the form of a long question that asks participants to select “Somewhat Disagree” at the end. Otherwise, the procedure for Pilot 2 was identical to Pilot 1.

#### S1.2.1 Pilot 2 Results:

In Pilot 2, we were most concerned with improving our experimental procedure. First, we examined whether participants believed they were speaking to real people during the chat portion of the experiment. We found that 183/199 (92%) participants believed that they were actually speaking to other participants during the experiment, which is similar to our results from Pilot 1. We then wanted to see whether we had increased the number of participants who believed they were watching a real person’s screen recording. With the changes to the procedure, we increased the believability from 52% to 70%. We found that 140/199 of participants believed that they were watching a real person’s screen recording.

## S1.3.0 Prolific Pilots

### *S1.3.1 Pilot Experiment 3*

In Pilot Experiment 3, we tested our procedure for Experiment 1 – Minimal Groups Design. We recruited 160 participants from the U.S. on Prolific. Of those participants, 149/160 were successfully matched into chat groups, which allowed them to start the experiment. Thus, the rest of the pilot analyses are conducted on those 149 participants who completed the experiment. We found that 96% of the participants passed the attention check, which is much higher than the population of NYU undergraduates from our previous pilots. Due to this massive increase in attention, we reduced the number of participants we are registering to collect for each experiment to N = 600, still above the 520 that were deemed necessary in our power analysis.

#### S1.3.2 Pilot 3 Results:

Among these participants, we found that 65% believed they were talking to real prolific workers, which is significantly greater than chance. We think the emergence of AI models like chatGPT and general skepticism from professional survey takers may be attributing to low believability. We also found that only 58% of participants believed they were watching a real screen recording, which was not significantly different from chance. Participants reported qualitatively that they were skeptical because they had not had to download any special software for the experiment.

### *S1.3.3 Pilot Experiment 4:*

We strove to improve these numbers in Pilot 4, where we tested our procedure for Experiment 2 - Natural Groups Design. Thus, we made several changes to our procedure First, we eliminated the screen recording from the paradigm. Participants are now told what another participant chose after a bogus waiting period. Second, we now required participants to sign up for our experiment in advance of their participation. In part, this was practically necessary, as Experiment 2 requires 2 Democrats and 2 Republicans to take the survey at the same time. Initial tests of the paradigm revealed that the likelihood of having the correct number and political orientation of participants by chance was extremely low, with very few participants being successfully matched into groups.

Thus, in our Pilot Experiment 4, we recruited 120 participants from the U.S. via Prolific. Specifically, we recruited 60 Democrats and 60 Republicans. Participants signed up in the morning of the experiment, and then received a survey link directly from the researchers at their scheduled participation time via the Prolific chat portal. Of the 120 participants who signed up, 51 participants entered the chat within 10 minutes of their assigned time, and of those, 44 participants were matched in groups and completed the experiment. Thus, our pilot sample for Experiment 2 is smaller than our pilot sample for Experiment 1, but this is mostly due to financial constraints from paying participants who were not successfully matched and were therefore unable to complete the experiment. We also found that only 1 participant failed the attention check, again much higher than the NYU undergraduate sample. This reinforces our decision to reduce our sample size to N=600 for our registered experiments.

#### S1.3.4 Pilot 4 Results:

In Pilot Experiment 4, we find much better results for believability. Compared to 65% from Pilot Experiment 3, 79.5% of participants believed they were talking to other Prolific workers during the chat section, significantly better than chance. Furthermore, 71.9% of participants now also believed that they had been informed of another real person’s real decision, also significantly better than chance. Thus, we believe that the additional aspect of scheduling for the experiment as well as simplifying the procedure has increased believability.

Regarding whether people make the immoral decision when given the choice, results are mixed. In our pilot of Experiment 1, we found that only 20/37 people in Condition 1 (the self condition) chose to assign tasks, the selfish choice. Of those 20, everyone chose to complete the green task. In our pilot of Experiment 2, we found that 8/12 people decided to assign tasks, and that all 8 of those participants assigned themselves the green task. Although it is a deviation from the original paper, this increase in altruistic behavior (or at the very least, fair/neutral behavior) may reflect a genuine change in demographics in the population.

**Open Practices Statement:** Proposed analysis scripts can be found at the following link: <https://github.com/clairerobertson/moral_hypocrisy_replication>. The data for this study will be deidentified and publicly accessible on OSF after data collection. This work is a preregistered replication.

1. The authors thank P. Valdesolo for providing more information on the procedure in personal correspondence. [↑](#footnote-ref-1)
2. Thank you to Reviewer 2 who sent us R code for these power simulations. [↑](#footnote-ref-2)
3. We encountered several challenges recruiting a nationally representative sample. Details and full demographic breakdown the sample are included in Supplement. First, because this study involved a live-chat, participants were recruited using a sign-up form, and instructed to return at a designated time to participate. However, many participants who signed up did not actually participate at the assigned time. Second, our Full details of recruitment and attrition are included in the Supplement XXX. [↑](#footnote-ref-3)