Supplementary Materials

Contents

A Pilot Studies

A.1 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.

A.1.1 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.

A.1.2 Pilot Experiment 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.

B Complier Average Causal Effects

In our preregistration, we preregisted a compliance average causal effect (CACE) analysis. We did this in order to adjust for a confound in the original study design whereby participants in the "Self" condition were given a choice as to whether they wanted to make an altruistic decision or selfish decision, whereas participants in all other conditions observed a selfish decision from another person. In the original paper, the authors excluded the participants who behaved altruistically. However, scholars caution against subsetting or excluding participants based on task choices (Lachin, 2000). However, we failed to meet several of the specific assumptions of the CACE analysis, and thus the results were not meaningful nor statistically interpretable.

CACE analyses rely on instrumental variables, and must meet five assumptions, as laid out in Connell (2010). In our experiment, we do not meet all five assumptions.

The first assumption states that outcomes for each participant are independent of one another, or the Stable Unit Treatment Value Assumption (SUTVA) (Connell, 2010). We do meet this criteria. Second, we must there is a monotonic relationship between treatment assignment and treatment receipt, meaning that no participants who were assigned to treatment had reduced likelihood of receiving treatment. We do not meet this assumption due to the lack of an interpretable CACE control condition. As quoted in Connell (2010) "The core insight behind CACE analysis is that we can arrive at an unbiased estimate of the difference in outcomes for compliers in the intervention group with those who would have engaged with treatment in the control group." We, however, had an inverse situation due to the structure of our study. The "outcome" of interest in our study was judgements of fairness after engaging in or witnessing a moral transgression. Our "controls" were those who had witnessed a moral transgression and our "treatment: group contained both compliers (those who had behaved immorally) and noncom pliers (those who had behaved altruistically). Thus, we already know the outcome for compliers in our control group, because the structure of the study required that they "comply" with the treatment. This instead means that the *only* participants who have the choice not to comply are those in our treatment condition. Third, CACE assumes that the compliance rate is not zero in the treatment condition, which we meet. Fourth, we assume that participants are randomly assigned to treatment, which we meet, However, we do not meet the second part of this fourth assumption, stating that the proportion of compliers across conditions should be the same. Again, this is because our "control condition" contains *only* compliers, not a random mix of potential compliers and non-takers. Fifth, we assume that random assignment to the treatment group does not affect the outcomes of individuals who do not comply with treatment, which we do not meet. It is indeed the non-compliers (altruists) that are different from the rest of our sample due to the fact that they behaved altruistically, which they were only able to do because of their random assignment.

We also lack an interpretable instrumental variable. EXPAND IF NEEDED.

Thus, our paradigm is an inappropriate fit for the CACE analysis, and we fail to meet three out of five for the CACE analysis. Attempts to estimate the CACE result instead estimations of outcomes (fairness ratings) for people in the control conditions (other, ingroup, outgroup) for if they would have made an altruistic decision, which is impossible given our paradigm.

We have included those estimations below, again for transparency with our preregistration, but we want to emphasize that they are are *both statistically and theoretically uninterpretable*.

INCLUDE TABLE HERE.

n a CACE analysis, the effect of a treatment is estimated based on those in the treatment group who complied, those in the treatment group who did not, and those who were never offered treatment. In our case, we were interested in how people who behaved altruistically in the self condition would have behaved in the other conditions.

C Recruitment and Demographic Information

C.1 Study 1 Recruitment and Demographic Information

In order to closely replicate the induction from Valdesolo and DeSteno (2007), we included a live chat in our study to induce participants to feel that they were participating in a study with real people. In order to ensure that sufficient participants were online at a given time to create a group of four as required by the study design, we recruited participants on Prolific in two parts. In Part 1, participants signed up to participate in our study at a given time later that day. They were then sent a message thanking them for tier interest, and letting them know that they would be receiving the link to Part 2 of the study (the real study) via Prolific's chat function at the time listed in the sign up study. They were then all sent a link to Part 2 at the designated time, asking them to begin the study within 10 minutes. Participants could thus sign up at any point in the day, but would all participate at the same time.

This ensured that the greatest number of participants would be matched successfully into groups to complete the study, but it also resulted in (a) significant attrition from Part 1 to Part 2 of the study, and (b) participant quotas that were not filled by the time the study was scheduled to be launched. In Study 1, we attempted to recruit 1100 nationally representative participants in 3 batches (n = 300, n = 400, n = 400). 937 participants signed up in Part 1, but only 584 participants joined Part 2 on time and were able to be matched with three other participants to complete the study.

We preregistered a sample size of 600. Because Prolific only allows nationally representative samples to be collected in batches of 300-500 participants, we decided to recruit an additional 80 participants as a standard sample to achieve a sample size of at least 600. Because this is a slight deviation from our preregistered plan, we conducted a robustness check excluding the non-representative sample, and found that they do not change our results SEE SUPPLEMENT XXXX.

Demographic information for Study 1 can be found in Table 1.

Variable	*n*
Ethnicity	
American Indian or Alaskan Native	3
Asian	28
Black or African American	77
Hispanic, Latino or Spanish Origin	20
Middle Eastern or North African	2
Native Hawaiian or Other Pacific Islander	1
White	421
Some other race, ethnicity or origin	2
Prefer not to say	4
Gender	
Man	295
Woman	275
Nonbinary	9
Something not listed here	5
Prefer not to say	4

Table 1: Study 1 Demographic Information

C.2 Study 2 Recruitment and Demographic Information

Variable	*n*
Ethnicity	
American Indian or Alaskan Native	0
Asian	0
Black or African American	0
Hispanic, Latino or Spanish Origin	0
Middle Eastern or North African	0
Native Hawaiian or Other Pacific Islander	1
White	421
Some other race, ethnicity or origin	2
Prefer not to say	4
Gender	
Man	295
Woman	275
Nonbinary	9
Something not listed here	5
Prefer not to say	4

Table 2: Study 2 Demographic Information NOT COMPLETE

D Study 1 Robustness Checks

D.1 Differences in Overestimators and Underestimators

No differences for over / under condition, and no differences in any of the pairwise tests

ADD HERE: ANOVA, PAIRWISE TESTS AND EQUIVALENCE TESTING RESULTS

D.2 Collective Identification and Fairness Ratings

ADD HERE: GRAPH WITH OUTGROUP IDENTIFIERS INCLUDED.

D.3 Intent to Treat Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	111.57	37.19	16.55	0.0000
Contrast 1 - Self vs. Other	1	91.25	91.25	40.61	0.0000
Contrast 2 - Ingroup vs. Outgroup	1	16.77	16.77	7.46	0.0065
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	3.56	3.56	1.58	0.2086
Risiduals	581	1305.57	2.25		

Table 3: INCOMPLETE: Intent-to-treat model n = XXX.

D.4 Manipulation Check

Because our study involved deception, it was critical that participants believe that they were (a) talking to real people during the chat phase and (b) that they were seeing a real person's choice behavior. We pretested our paradigm in several pilot studies, and in our final sample for Study 1, we found that XXXX% of participants believed they were talking to a real person during the chat phase of the experiment, and XXX % of participants in Conditions 2, 3, and 4 believed that another participant really was assigning tasks. As a robustness check, we found that our results were robust even when excluding those who failed both manipulation checks.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	15.83	5.28	2.54	0.0568
Contrast 1 - Self vs. Other	1	0.23	0.23	0.11	0.7419
Contrast 2 - Ingroup vs. Outgroup	1	12.54	12.54	6.03	0.0146
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	3.06	3.06	1.47	0.2263
Risiduals	305	634.06	2.08		

Table 4: INCOMPLETEOnly those who passed the manipulation check, *n* = XXX

E Study 1 Exploratory Analyses

E.1 Political Ideology and Extremity

ADD HERE - IDEOLOGY AND EXTREMITY FINDINGS.

E.2 High Identifiers Only

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	37.83	12.61	4.55	0.0045
Contrast 1 - Self vs. Other	1	9.66	9.66	3.49	0.0639
Contrast 2 - Ingroup vs. Outgroup	1	28.15	28.15	10.16	0.0018
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	0.03	0.03	0.01	0.9241
Risiduals	138	382.26	2.77		

Table 5: Only partiicpants who scored greater than 2 on Collective Identification, *n* = 142

F Study 2 Robustness Checks

F.1 Intent to Treat Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	111.57	37.19	16.55	0.0000
Contrast 1 - Self vs. Other	1	91.25	91.25	40.61	0.0000
Contrast 2 - Ingroup vs. Outgroup	1	16.77	16.77	7.46	0.0065
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	3.56	3.56	1.58	0.2086
Risiduals	581	1305.57	2.25		

Table 6: Study 2: Intent-to-treat model, *n* = 585.

F.2 Differences in Overestimators and Underestimators

ADD HERE: ANOVA RESULTS, PAIRWISE T TESTS AND EQUIVALENCE TESTS.

F.3 Politically Mismatched Participants Excluded

Participants were assigned roles (either Democrats or Republicans) based on the ideology they reported 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 there were 5 participants who had reported that the were a Democrat on Prolific but responded that they were conservative in the survey, and 9 participants who reported that they were a Republican on Prolific, but responded that they were liberal in the survey.

We also asked participants how much they identified with their political ingroups or outgroups in the survey, and found that there were 11 Democrats who reported identifying more with Republicans than Democrats, and 25 Republicans who reported identifying more with Democrats than Republicans. Results do not significantly change when these participants (*n* = 46) are excluded. Results from the contrast models are listed in Table 7, and results from the linear model examining the interaction effect between condition (Ingroup vs. Outgroup) and Collective Identification was not significant $\beta = 0.086$, *t*(244) = 0.74, *p* = 0.46.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	29.16	9.72	4.75	0.0028
Contrast 1 - Self vs. Other	1	2.38	2.38	1.16	0.2816
Contrast 2 - Ingroup vs. Outgroup	1	21.39	21.39	10.46	0.0013
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	5.39	5.39	2.64	0.1052
Risiduals	482	985.87	2.05		

Table 7: Politically mismatched participants excluded, *n* = 486.

F.4 Repeat Participants and Groups Excluded

Due to an error on Prolific, there were *n* = 13 participants who participated in our study more than once. We exclude their repeated participations from analyses, but because this study involves both deception and participant interaction, it was possible that the repeat subjects could have revealed the purpose of the study to other participants during the chat phase. After examining the chat logs, we found repeat participants did not reveal the purpose of the study during the chats. Nonetheless, we ran a robustness check excluding all participants who chatted with a repeat participant *n* = 45. Overall, we found that our results were robust even when excluding these participants.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	23.66	7.89	3.82	0.0101
Contrast 1 - Self vs. Other	1	3.09	3.09	1.50	0.2219
Contrast 2 - Ingroup vs. Outgroup	1	18.32	18.32	8.87	0.0030
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	2.25	2.25	1.09	0.2973
Risiduals	499	1031.32	2.07		

Table 8: Results excluding those who had chatted with a repeat participant, *n* = 503

F.5 Manipulation Check

Because our study involved deception, it was critical that participants believe that they were (a) talking to real people during the chat phase and (b) that they were seeing a real person's choice behavior. We pretested our paradigm in several pilot studies, and in our final sample for Study 2, we found that 73.97% of participants believed they were talking to a real person during the chat phase of the experiment, and 74.48 % of participants in Conditions 2, 3, and 4 believed that another participant really was assigning tasks. As a robustness check, we found that our results were robust even when excluding those who failed both manipulation checks.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	15.83	5.28	2.54	0.0568
Contrast 1 - Self vs. Other	1	0.23	0.23	0.11	0.7419
Contrast 2 - Ingroup vs. Outgroup	1	12.54	12.54	6.03	0.0146
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	3.06	3.06	1.47	0.2263
Risiduals	305	634.06	2.08		

Table 9: Only those who passed the manipulation check, *n* = 309.

G Study 2 Exploratory Analyses

G.1 Political Ideology and Extremity

G.2 Partisan Differences

We did not preregister any partisan differences, but we also wanted to include an exploratory analysis comparing Democrats and Republicans. We find evidence of outgroup derogation, but not moral hypocrisy in Democrats, whereas we find evidence of both outgroup derogation and moral hypocrisy in Republicans.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	11.08	3.69	1.96	0.1198
Contrast 1 - Self vs. Other	1	3.14	3.14	1.67	0.1976
Contrast 2 - Ingroup vs. Outgroup	1	7.88	7.88	4.19	0.0417
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	0.07	0.07	0.04	0.8505
Risiduals	264	496.63	1.88		

Table 10: Democrats only, $n = 268$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	3	39.75	13.25	6.12	0.0005
Contrast 1 - Self vs. Other	1	20.35	20.35	9.41	0.0024
Contrast 2 - Ingroup vs. Outgroup	1	10.76	10.76	4.97	0.0266
Contrast 3 - Self/Ingroup vs. Other/Outgroup	1	8.64	8.64	3.99	0.0467
Risiduals	260	562.40	2.16		

Table 11: Republicans only, $n = 264$.