# Investigating the relationship between self-perceived moral superiority and moral behavior using economic games

Ben M. Tappin [a, 1, *] & Ryan T. McKay [a, 2]

[a] ARC Centre of Excellence in Cognition and its Disorders, Department of Psychology, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
Institutional telephone number: +44 (0) 1784 434455 (Royal Holloway)

Email: [1] Ben.Tappin.2015@live.rhul.ac.uk; [2] Ryan.McKay@rhul.ac.uk

* Corresponding author

**This is a pre-proof version of the article in press at *Social Psychological and Personality Science*.**

Most people report that they are superior to the average person on various moral traits. The psychological causes and social consequences of this phenomenon have received considerable empirical attention. The behavioral correlates of self-perceived moral superiority, however, remain unknown. We present the results of two preregistered studies (Study 1, N=827; Study 2, N=825) in which we indirectly assessed participants' self-perceived moral superiority, and used two incentivized economic games to measure their engagement in moral behavior. Across studies, self-perceived moral superiority was unrelated to trust in others and to trustworthiness, as measured by the Trust Game; and unrelated to fairness, as measured by the Dictator Game. This pattern of findings was robust to a range of analyses, and, in both studies, Bayesian analyses indicated moderate support for the null over the alternative hypotheses. We interpret and discuss these findings, and highlight interesting avenues for future research on this topic.

Self-perceptions of moral superiority appear robust and relatively widespread. In numerous studies, majorities of people rate themselves as fairer, more trustworthy, more honest—more *moral*—than the average person (Epley & Dunning, 2000; Fetchenhauer & Dunning, 2006; Klein & Epley, 2016, in press; Tappin & McKay, 2017; Van Lange & Sedikides, 1998). Under the broader phenomenon of "self-enhancement" (Alicke & Sedikides, 2011), past work has investigated (i) psychological explanations for (Sedikides et al., 2014; Tappin & McKay, 2017; Van Lange & Sedikides, 1998), and (ii) interpersonal consequences of (Barranti et al., 2016; Heck & Krueger, 2016) self-perceived moral superiority. There is a conspicuous lack of evidence, however, for how these perceptions relate to engagement in behaviors commonly considered moral—such as freely helping others, or reciprocating trust. In the present article, we report an initial investigation of this relationship.

## Self-perceived moral superiority and engagement in moral behavior

There exists much debate over whether the prevalence of self-superiority phenomena is best explained by motivational or non-motivational processes (Brown, 2012; Chambers & Windschitl, 2004; Taylor & Brown, 1988). This offers a useful framework for speculating on how self-perceived moral superiority may relate to engagement in moral behavior.

Consider people who perceive themselves to be *strongly* morally superior to the average person. As a function of their strong sense of righteousness relative to other people, these individuals may be motivated to behave in (moral) ways to protect this positive social comparison. According to various reviews, self-protection is a fundamental human motivation (Sedikides et al., 2015), and social comparison a common process by which people derive positive self-evaluation (Sedikides & Strube, 1997; Wills, 1981). Moral traits, moreover, are held in high regard (Van Lange & Sedikides, 1998), and morality appears to be central to notions of identity (Strohminger & Nichols, 2014, 2015). Individuals who possess a *weaker* sense of righteousness over the average person, then, may accordingly possess a relatively

weaker motivation to protect the (less positive) social comparison. This implies that self-perceived moral superiority may be positively associated with engagement in moral behavior.

Another motivational process that might predict a positive association is sensitivity to the charge of hypocrisy. Hypocrites are loathed—more so than people who are honest about their moral limitations (Jordan et al., 2017)—and especially so when the hypocrite considers themselves to be superior to others (Alicke et al., 2013). Heck and Krueger (2016) recently reported evidence that agents who made inaccurate claims of moral self-superiority received the strongest moral condemnation from observers; stronger, even, than agents who *accurately* reported being *less* moral than the average person. Put another way, observers punished people most when their self-reported moral superiority was shown to be false by their behavior. These findings imply added motivation for such people to behave morally, so to avoid harsh social censure. Consistent with this suggestion is evidence that individuals behave more prosocially after criticizing another person (Simpson et al., 2013).

Some non-motivational processes, on the other hand, may lead us to expect a negative association between self-perceived moral superiority and engagement in moral behavior. Fetchenhauer and Dunning (2009, 2010) provide evidence that individuals underestimate the moral goodness (specifically, trustworthiness) of other people due to an informational asymmetry in the social environment. If person A decides to trust person B, this occasionally results in surprising and costly betrayal by person B. In contrast, when person A decides *not* to trust person B, this necessarily *precludes* person A learning that person B was, in fact, trustworthy. The implication is that individuals learn asymmetrically about the trustworthiness of other people; an asymmetry which may underlie cynicism about the moral goodness of others more generally (Fetchenhauer & Dunning, 2010; Miller, 1999).

Such a mechanism could help explain the prevalence of self-perceived moral superiority. Specifically, because the lion’s share of the variance in self-perceived moral

superiority likely derives from variance in how people perceive the moral goodness of *others*, rather than themselves. There is relatively limited variance in the latter—people seem largely in agreement that they *themselves* are morally virtuous (for a brief review, see Tappin & McKay, 2017). Taking this in conjunction with evidence that—in interdependent contexts—individuals' moral behavior is conditional on whether they think *others* will behave in kind (Krueger & Acevedo, 2007) implies that greater cynicism—and, thus, greater self-perceived moral superiority—may be associated with *less* moral behavior.

## Overview

Given the uncertainty over how self-perceived moral superiority relates to engagement in moral behavior, we set out to investigate this relationship. Specifically, across two studies, we used canonical economic games as measures of moral behavior, and indirectly assessed how moral individuals perceived themselves to be relative to the average person.

## Methods

The preregistered protocols, analysis scripts and data for both studies are available on the Open Science Framework (OSF): https://osf.io/p42mp/. Because of their similarity, we present the methods and results of these studies together.

### Engagement in moral behavior

To measure engagement in moral behavior, we used two incentivized, one-shot, anonymous economic games (with no deception); the Trust Game (TG, Study 1) and Dictator Game (DG, Study 2). These games are typically taken as providing measures of trust in others and trustworthiness, and fairness[1], respectively (see below for descriptions of the games).

While a general prosocial preference is likely to underpin behavior in both the TG and DG (Peysakhovich et al., 2014), past work suggests that trusting behavior in the TG is distinct

[1] We refer to the DG as measuring "fairness" throughout, but note that giving in the DG is also consistent with altruism (Rand et al., 2016). In analyses, we find little difference in the results depending on how the DG measure is construed.

from giving in the DG (Brülhart & Usunier, 2012), and, indeed, a recent large investigation reported that the shared variance between trusting behavior in the TG, and behavior in the DG, was relatively modest at 12% (Peysakhovich et al., 2014). The relationship between DG behavior and trust*worthy* behavior in the TG was estimated to be somewhat higher—at 25% shared variance with behavior in the DG. In both cases, however, there was evidence of unique variance between the games. This suggests that inclusion of both the TG and DG provided us with three somewhat overlapping but distinct measures of behavior.

We used economic games to measure engagement in moral behavior because numerous studies indicate that people subjectively imbue choices in these games with moral weight. For example, recent evidence suggests that prosocial behavior in economic games is driven by an explicit preference for behaving morally (Capraro & Rand, 2017), and behaving prosocially in such games is consistently and strongly judged to be morally superior to behaving self-interestedly (Krueger & Acevedo, 2007; Krueger & DiDonato, 2010; Krueger et al., 2008). The inclusion of the TG and DG thus provided a straightforward decision environment with a recognizable "moral" behavior.

**Trust Game.** In our TG, participants are anonymously paired and assigned the role of either "Trustor" or "Trustee". Both participants are given $0.20 as a starting endowment, and the Trustor has the option to transfer any amount of their endowment to the Trustee (from $0.00 to $0.20 in increments of $0.01). Any amount they transfer is tripled on its way to the Trustee, and the Trustee is then able to decide how much, if any, of this tripled amount they would like to transfer back to the Trustor (from 0 to 100%). Since the Trustor takes a risk by sending money to the Trustee, their decision is usually taken as a measure of trust. The Trustee, on the other hand, has the option to reciprocate the trust placed in them by the Trustor, by sending some amount of money back to the Trustor. The Trustee decision is thus usually taken as measure of trustworthiness (e.g., Berg et al., 1995).

**Dictator Game.** In our DG, participants are anonymously paired and assigned the role of either "Dictator" or "Receiver". The Dictator is given $0.30 as a starting endowment, whereas the Receiver starts with nothing. The Dictator then has the option to transfer any amount of their endowment to the Receiver (from $0.00 to $0.30 in increments of $0.01). Since the Dictator's decision is unilateral, with no possibility of reciprocation (or punishment) from the Receiver, they have no financial incentive to share the money. As such, the Dictator's decision to share money is usually taken as a measure of fairness (more technically, *inequity aversion*, see Fehr & Schmidt, 1999).

**Self-perceived moral superiority**

To measure self-perceived moral superiority[2], we used a regression-based index of trait self-superiority developed and described in detail elsewhere (Heck & Krueger, 2015; Tappin & McKay, 2017). In brief, participants are asked to judge the extent to which 10 moral traits describe (i) themselves and (ii) the average person. They also rate (iii) the social desirability of the traits. The moral traits are presented in Table 1. Conventional measures of self-superiority typically compare how positive self-judgments are with respect to judgments of the average person. However, this overestimates the magnitude and frequency of people who harbor perceptions of self-superiority. The current measure accounts for this overestimation by estimating—and allowing the researcher to remove—a component of self-superiority that may be deemed "defensible" because of the uncertainty people face when making judgments of the average person. Below we describe the computational steps of the measure only (for more detail, see Heck & Krueger, 2015; Tappin & McKay, 2017).

[2] In both preregistrations, this construct is referred to as "self-righteousness". This was relabelled to "self-perceived moral superiority" during the review process for better linguistic and conceptual clarity. The measure is identical to that described in the preregistrations.

Table 1.

Positive and negative moral traits used in Studies 1 and 2

| Positive moral traits | Negative moral traits |
|---|---|
| Honest | Insincere |
| Trustworthy | Prejudiced |
| Fair | Disloyal |
| Respectful | Manipulative |
| Principled | Deceptive |

*Note.* We used the five positive and five negative moral traits from Tappin and McKay (2017).

***Step 1.*** We first estimate how similar each participant's moral self-judgments are to those of the average participant in the sample. To do so, we calculate the average self-judgment for each moral trait over all participants, and then regress these averages on the moral self-judgments made by each individual participant. This provides a moral "coefficient of similarity" (unstandardized slope, *b*) and intercept for each participant. Higher coefficient values indicate that the participant is more like the average participant in the sample. We then compute the mean moral coefficient of similarity and intercept across participants.

***Step 2.*** Next, we generate *inferred* moral self-judgments (I) by weighting participants empirically-observed moral judgments of the average person (O) by the mean coefficient of similarity and intercept, using the formula:

$$I = \frac{O}{mean\ coefficient\ of\ similarity} + mean\ intercept$$

Inferred self-judgments represent self-judgments an *ideal* judge would have made. That is, a judge who perceives how morally similar people are, and uses this information to weight their judgment of the average person to make a more accurate self-judgment. (The basic

rationale is this: the more similar people are—defined here by the mean *coefficient of similarity*—the less participants' self-judgments are expected to deviate from their judgments of the average person; see Heck & Krueger, 2015; Tappin & McKay, 2017). At this stage, then, each participant has four sets of judgments for the 10 moral traits. Their empirically observed self-judgments (S), judgments of the average person (O), and social desirability judgments (D), and the new inferred self-judgments (I) computed according to the preceding method.

***Step 3.*** In the final step, we regress S, O, and I on D judgments for each participant. This produces three unstandardized slopes per participant. These slopes express how well moral trait desirability predicts their (i) moral self-judgments ($b_{SD}$), (ii) judgments of the average person ($b_{OD}$), and (iii) inferred self-judgments ($b_{ID}$). In other words, $b_{SD}$ describes the positivity of participants' moral self-perception, $b_{OD}$ describes the positivity of participants' perception of the average person's morality, and $b_{ID}$ describes the positivity of the participants' moral self-perception presupposing they were an ideal judge.

The index of self-perceived moral superiority is computed as the difference between $b_{SD}$ and $b_{ID}$ (specifically, $b_{SD}$ - $b_{ID}$). This index represents self-perceived moral superiority, but is more conservative than conventional measures because it partitions out a "defensible" component of self-superiority (which is defined by the difference between $b_{ID}$ and $b_{OD}$)[3].

**Samples**

We sought to recruit 824 participants in each study, providing approximately N=412 in each role. Participants were recruited via Amazon's Mechanical Turk (Amir & Rand, 2012; Arechar et al., in press; Chandler & Shapiro, 2016; Rand, 2012). Sample sizes were determined via power analyses: Our smallest effect size of interest was $r$=.15, which we required N=343 to achieve 80% power ($\alpha$=.05) to detect in each of our three primary linear regression analyses

[3] We report correlations between the "defensible" component of self-perceived moral superiority and economic game behavior in the SI (section 5).

(Faul et al., 2009). We deliberately oversampled by approximately 20% to guard against power loss due to planned data exclusions. Sample sizes after data collection were: Study 1 N=827 (50.18% female, $M_{age}$=38.35 $SD_{age}$=12.97; Trustor N=413, Trustee N=414), Study 2 N=825 (55.27% female, $M_{age}$=37.50 $SD_{age}$=12.62; Dictator N=413, Receiver N=412).

**Procedure**

The procedure in both studies was substantively identical, and we recruited separate samples in each case (Study 1 participants were identified via their unique Mechanical Turk ID and blocked from participating in Study 2). All participants provided informed consent, before being assigned their role in their respective economic game (Study 1: TG, trustor or trustee, Study 2: DG, dictator or receiver, role assignments were counterbalanced). All participants then completed (i) the trait judgment task, and (ii) the economic game (counterbalanced), except for those assigned the role of receiver in the DG. These participants always completed the DG first, and then completed an unrelated task (receivers are entirely passive and so collecting their trait judgments was unnecessary).

In the trait judgment task, participants were presented with the list of 10 moral traits alongside 20 additional, nonmoral filler traits (inclusion of the nonmoral traits allowed us to replicate the primary results reported by Tappin & McKay, 2017; see SI section 6). Participants were asked to judge (i) the extent to which each trait described themselves, (ii) the extent to which each trait described the average person, and (iii) the social desirability of each trait. Participants rated all 30 traits according to either (i), (ii), or (iii), before moving onto the next set of ratings, and the order of these three sets of judgments was counterbalanced across participants. The presentation order of the traits themselves was randomized in each rating set and for each participant. Rating judgments for the self and the average person were provided on a seven-point scale, ranging from 1 (*Not at all*) to 7 (*Very much so*). Social desirability

judgments were also provided on a seven-point scale, ranging from -3 (*Very undesirable*) to +3 (*Very desirable*).

In the economic games, participants read instructions and completed three comprehension questions assessing their understanding of the payoff structure. Failure to answer all three comprehension questions correctly after two attempts resulted in participants being prevented from completing the survey. After these questions, we revealed which role the participant had been assigned, and they made their decision. We informed them that pairs of decisions would be combined and their bonus calculated and awarded after the survey had concluded (which was true). In addition to bonuses, all participants received a base fee of $0.50 for taking part. At the end of the survey, participants completed simple demographic questions, provided feedback on their experience, and were asked whether they had previously taken part in a similar decision task.

**Results**

All analyses were conducted in the R environment (R Core Team, 2016). Only dictators are used in Study 2 analyses. Table 2 displays descriptive statistics.

Table 2.

Descriptive statistics from Studies 1 and 2

| | Study 1 (TG) | | | | Study 2 (DG) | | | |
|---|---|---|---|---|---|---|---|---|
| | Slope (b) | | Intercept | | Slope (b) | | Intercept | |
| | M | SD | M | SD | M | SD | M | SD |
| Components | | | | | | | | |
| $R_{SD}$ | 0.74 | 0.27 | 0.98 | 1.15 | 0.76 | 0.26 | 0.90 | 1.09 |
| $R_{OD}$ | 0.19 | 0.34 | 3.13 | 1.47 | 0.18 | 0.36 | 3.15 | 1.50 |
| $R_{ID}$ | 0.22 | 0.41 | 4.38 | 1.75 | 0.21 | 0.41 | 4.14 | 1.72 |

| | M | SD | M | SD |
|---|---|---|---|---|
| Index of SPMS | 0.52 | 0.41 | 0.55 | 0.44 |
| Transfer amount | | | | |
| Trustors (c) | 13.38 | 7.28 | - | - |
| Trustees (%) | 35.24 | 24.44 | - | - |
| Dictators (c) | - | - | 10.39 | 6.96 |

*Note*. Components are within-participant regressions involved in computing the index of self-perceived moral superiority, according to the procedure outlined in the methods section. TG = Trust Game, DG = Dictator Game; R = regression; S = self-judgments; D = desirability judgments; O = other (average person) judgments; I = inferred-self judgments; SPMS = self-perceived moral superiority (i.e., $b_{SD} - b_{ID}$); M = mean; c = cents. Study 1 N = 736, Study 2 N = 369.

**Data exclusions**

All data exclusions were preregistered. Before computing the self-perceived moral superiority index, we excluded responses that contained duplicate IP addresses (Study 1: n=8, 0.97%, Study 2: n=2, 0.48%) and/or one or more failed attention checks (there were three embedded in the trait judgment task) (Study 1: n=28, 3.39%, Study 2: n=22, 5.33%). We then proceeded to compute the index as outlined in Steps 1-3 in the methods section. During Step 1, those participants who responded uniformly on moral self-judgments were excluded (Study 1: n=0, 0%, Study 2: n=4, 0.97%), because the regression analyses in this step require at least *some* variation. During Step 3, for the same reason, we additionally excluded participants who responded uniformly on moral judgments of the average person (Study 1: n=54, 6.53%, Study

2: n=19, 4.60%), and/or social desirability judgments (Study 1: n=1, 0.12%, Study 2: n=4, 0.97%). Sample sizes for the primary analyses were thus, Study 1: Trustors N=369, Trustees N=367, Study 2: Dictators N=369.

**Self-perceived moral superiority and trust in others**

**Preregistered analyses.** We first regressed trustor decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to transfer amount, model summary: F (1, 367) = 0.14, p=.706, R=.02 [predictor summary: b=-0.34, se=0.89, t=-0.38]. Because the decision data were non-normally distributed, we also conducted a Spearman's rank correlation with the same two variables. The results mirrored the parametric analysis: $r_s$=-.05, p=.326. Magnitude of self-perceived moral superiority was not meaningfully associated with trusting behavior in the TG.

**Exploratory analyses.** We conducted several exploratory analyses to test the robustness of this conclusion. First, we dichotomized the trustor decisions by assigning them a value of 1 if they were greater than the median transfer amount of 15c, and a value of 0 if they were equal to or less than this amount. A total of 179 (48.51%) participants transferred greater than the median amount of 15c. A binary logistic regression predicting the probability of an above median transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: Odds Ratio (OR)=0.90, 95% CI [0.55, 1.46], p=.669 (Figure 1). That is, self-perceived moral superiority was not meaningfully associated with the probability of transferring greater than the median transfer amount. For all DVs, we also explored whether prior experience with the games was masking an association between self-perceived moral superiority and decision behavior in our sample (it wasn't) (cf. Chandler et al., 2015; see SI section 2 for these analyses).

Given these results, we sought to quantify the relative strength of evidence in favor of the null hypothesis. We conducted a Kendall's tau Bayesian correlation analysis using JASP

software (JASP Team, 2017). Under a uniformly distributed prior, we obtained a Bayes Factor (BF) of 8.23 in favor of the null hypothesis. That is, the BF indicated moderate support for the null over the alternative hypothesis. The BF in favor of the null remained moderate-to-strong over a wide range of priors (see SI section 1). The results of the exploratory analyses support those of the preregistered analyses.

**Self-perceived moral superiority and trustworthiness**

**Preregistered analyses.** As before, we began by regressing trustee decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to back-transfer amount: F (1, 365) = 0.04, p=.851, R=.01 [b=0.60, se=3.17, t=0.19]. Because the decision data were again non-normally distributed, we followed with a Spearman's rank correlation. The results mirrored the parametric analysis: $r_s$=-.004, p=.931. Magnitude of self-perceived moral superiority was not meaningfully associated with trustworthiness behavior in the TG.

**Exploratory analyses.** Once again, we conducted several exploratory analyses to investigate the robustness of this conclusion. We first dichotomized the trustee decisions by assigning them a value of 1 if they were greater than the median back-transfer amount of 50%, and a value of 0 if they were less than this amount. A total of 55 (14.99%) participants back-transferred greater than the median amount of 50%. A binary logistic regression predicting the probability of an above median back-transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: OR=0.64 [0.31, 1.32], p=.233 (Figure 1). That is, self-perceived moral superiority was not meaningfully associated with the probability of an above median back-transfer. As before, a Kendall's tau Bayesian correlation analysis conducted in JASP (uniformly distributed prior) returned a BF of 14.58 in favor of the null hypothesis. That is, the BF indicated strong support for the null over the alternative hypothesis. The BF remained

moderate-to-strong over a wide range of priors (see SI section 1). The results of the exploratory analyses thus support those of the preregistered analyses.

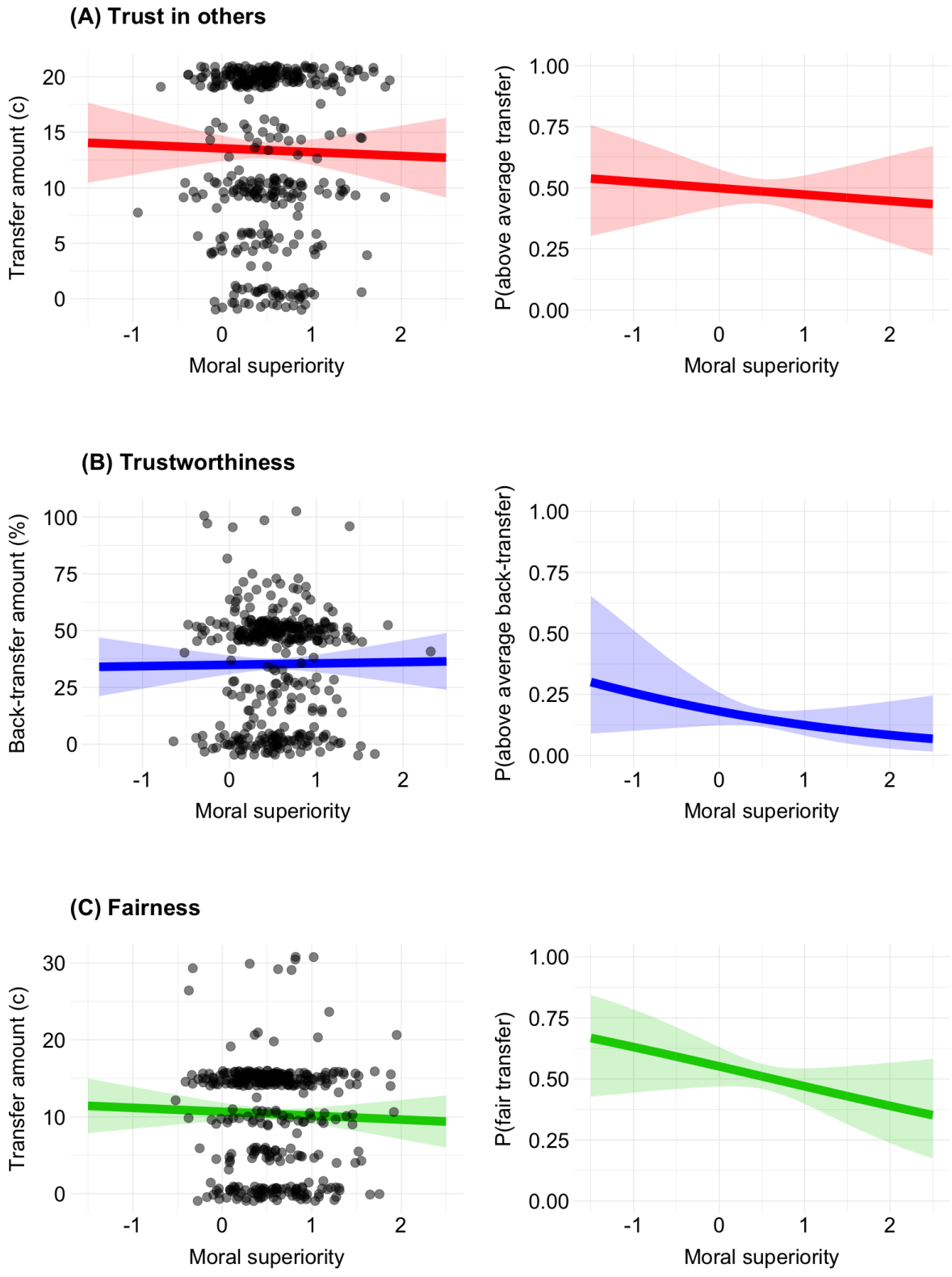


**Figure 1 | Relationship between self-perceived moral superiority and transfer amounts in studies 1 (A, B) and 2 (C). Scatter points are raw data with slight jitter for visibility and the shaded regions denote 95% confidence intervals. (A)** Left panel: Preregistered analysis

regressing trustor transfer amount on self-perceived moral superiority (b = -0.34, se = 0.89). Right panel: Exploratory binary logistic regression analysis of the probability that trustor transfer was greater than the median transfer amount (15c), based on self-perceived moral superiority scores (OR = 0.90 [0.55, 1.46]). N = 369. **(B)** Left panel: Preregistered analysis regressing trustee back-transfer amount on self-perceived moral superiority (b = 0.60, se = 3.17). Right panel: Exploratory binary logistic regression analysis of the probability that trustee back-transfer was greater than the median back-transfer amount (50%), based on self-perceived moral superiority scores (OR = 0.64 [0.31, 1.32]). N = 367. **(C)** Left panel: Preregistered analysis regressing dictator transfer amount on self-perceived moral superiority (b = -0.62, se = 0.83). Right panel: Exploratory binary logistic regression analysis of the probability that dictator transfer was fair (15c), based on self-perceived moral superiority scores (OR = 0.72 [0.45, 1.15]). N = 369.

### Self-perceived moral superiority and fairness

**Preregistered analyses.** We began by regressing dictator decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to transfer amount: F (1, 367) = 0.57, p=.452, R=.04 [b=-0.62, se=0.83, t=-0.75]. Because the decision data were non-normally distributed, we conducted a Spearman's rank correlation with the same two variables. The results mirrored the parametric analysis: $r_s$=-.05, p=.345. We quantified the relative strength of evidence in favor of the null by conducting a Bayesian correlation analysis in JASP. We preregistered our intention to conduct a *Pearson's rho* Bayesian correlation, but, given the severe non-normality of the decision data, a Kendall's tau Bayesian correlation is more appropriate. For transparency, we report both. The $BF_{rho}$ was 11.57, and $BF_{tau}$ was 8.38 in favor of the null hypothesis (uniformly distributed priors). Both indicated moderate-to-strong support for the null over the alternative hypothesis. In SI section 1, we report $BF_{tau}$ over a wide range of priors (it remained moderate-to-strong in favor of the null).

Next, to account for the fact that transfer amounts of greater than 15c—that is, greater than half the dictator's endowment—are technically "unfair" (Fehr & Schmidt, 1999), rather reflecting altruism or "hyper-fairness" (Henrich et al., 2006; Rand et al., 2016), we repeated the above analyses with a truncated sample of dictators—excluding those who transferred

greater than 15c ($N_{excluded}$=15, 4.07%). The truncated analyses thus tested whether self-perceived moral superiority was associated with fairness behavior, where unfair behavior was defined as inequity in favor of oneself (i.e., the dictator). The pattern of results was the same as in the full sample, regression: F (1, 352) = 0.87, p=.351, R=.05 [b=-0.73, se=0.78, t=-0.93], Spearman's rank correlation: $r_s$=-.06, p=.235, Bayesian correlation: $BF_{rho}$=9.75 and $BF_{tau}$=5.88 in favor of the null (uniform priors; $BF_{tau}$ robust over a range of priors, see SI section 1). Magnitude of self-perceived moral superiority was not meaningfully associated with fairness behavior in the DG.

**Exploratory analyses.** To check robustness, we dichotomized the dictator decisions by assigning them a value of 1 if they were equal to 15c, and a value of 0 if they were greater than or less than this amount. Fairness was thus strictly defined as rejection of inequity in favor of either oneself (dictator) *or* the other person (receiver). A total of 187 (50.68%) participants split the money fairly, transferring exactly 15c. A binary logistic regression predicting the probability of fair transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: OR=0.72 [0.45, 1.15], p=.174 (Figure 1). That is, self-perceived moral superiority did not meaningfully predict the probability of a fair transfer. The results of the exploratory analyses are consistent with those of the preregistered analyses.

## Discussion

We investigated how self-perceived moral superiority related to behavior in two canonical economic games—the Trust Game (TG) and Dictator Game (DG). Across two studies, self-perceived moral superiority was not associated with magnitude of trust in others, trustworthiness, or fairness, as these behaviors are measured in the games. This pattern of results was robust to a variety of analyses, and, for each of the three dependent variables, Bayesian analyses indicated relatively strong support for the null vs. alternative hypothesis.

The findings are inconsistent with our hypotheses: that self-perceived moral superiority would be associated with (i) more, or with (ii) less, moral behavior. Whereas some evidence suggests that perceptions of *non*moral self-superiority are associated with (Blanton et al., 1999; Heck & Krueger, 2015), and possibly facilitate (O'Mara & Gaertner, 2017) behavioral performance, we found that self-perceived moral superiority was not associated with behavior in canonical economic games—in which moral motivation appears reliably engaged (Capraro & Rand, 2017), and where morally superior decisions are readily discerned (Krueger & Acevedo, 2007; Krueger & DiDonato, 2010; Krueger et al., 2008).

Why was self-perceived moral superiority unrelated to behavior in the games? One explanation is that our measure was *domain general*. That is, participants provided judgments for a range of moral traits, which fed into a single score indexing their self-perceived moral superiority. It is possible that superiority perceived on specific moral traits *is* associated with behavior representative of those traits, but that our domain general measure obscured these relationships. We examined this possibility by computing raw difference scores between participants' self-judgments and their judgments of the average person for the traits "trustworthy" and "fair" only, and correlating these scores with trustee decisions, and dictator decisions, respectively (SI section 3). These coefficients were also trivial in size ($|r_s| < .03$)—suggesting that the domain-generality of our measure does not account for the current pattern of results.

An interesting and related question is whether individuals' moral *self*-perception—not their perceived *superiority* over others—was associated with absolute magnitude of monetary transfer in the games. Exploratory correlations suggested a small but consistently positive association between moral self-perception ($b_{SD}$) and transfer amount across dependent variables; trust in others ($r_s$=.12), trustworthiness ($r_s$=.15), and fairness ($r_s$=.06). We observed some evidence for self-knowledge—those people who had a more positive view of their own

morality tended to transfer more money to their partners. This is consistent with prior evidence that self-perceptions are at least somewhat diagnostic of behavior/reality (Epley & Dunning, 2000; Vazire & Carlson, 2010), and that self-reported traits correlate with prosociality in economic games (Hillbig et al., 2013). This raises the question of what role moral judgments of the average person had in participants' behavior.

It is plausible that the magnitude of self-perceived moral superiority is driven primarily by variance in how people view the morality of *other* people, not themselves (cf. Tappin & McKay, 2017), and that greater moral cynicism about others is associated with lower engagement in certain types of moral behavior (Krueger & Acevedo, 2007). This provides one explanation for why the above positive associations between moral self-perception and behavior did not emerge for self-perceived moral superiority. Specifically, because they were cancelled out by the cynicism disproportionately driving the latter.

We subjected this speculation to the data. First, comparing the shared variance between self-perceived moral superiority scores and both (i) moral self-perceptions ($b_{SD}$), and (ii) perceptions of the average person's morality ($b_{OD}$), revealed that the latter explained, on average, 64% variance in the scores, whereas the former accounted for less than a quarter of this amount (SI section 4). Second, perceptions of the average person's morality were weakly but consistently positively related to transfer amount across dependent variables; trust in others ($r_s$=.11), trustworthiness ($r_s$=.12), and fairness ($r_s$=.08). In other words, self-perceived moral superiority was mainly driven by how individuals viewed the morality of other people, not themselves, and greater moral cynicism about these others tended to be associated with lower monetary transfers. This supports our speculation on both counts, and is consistent with two areas of prior work: the first, that observers interpret expressions of self-superiority as condemnation of others, rather than egregious self-flattery (Van Damme et al., 2016; Van

Damme et al., 2017), and, the second, that individuals condition their behavior in these games on whether they think *others* will behave in kind (Krueger & Acevedo, 2007).

Based on this, we suggest that, despite the robust observation that most people rate themselves as morally superior to the average person, this phenomenon has limited predictive validity due to the seemingly opposed behavioral influences of self- and other-perception that comprise its measurement. That said, we note there is mixed evidence over whether economic games are valid analogues of behavior in the real world (Benz & Meier, 2008; Fehr & Leibbrandt, 2011; Franzen & Pointner, 2013; Galizzi & Navarro-Martinez, 2017). It is thus reasonable to ask whether our results would generalize to more ecologically valid cases of moral behavior. This represents an interesting avenue for future research. Furthermore, there is evidence that East Asian samples do not report self-superiority perceptions to the same extent as Western samples (Heine & Hamamura, 2007); indicating our results may differ along these specific cultural lines.

We do expect, however, that our results will be robust to variations in the economic game environment—in particular, changes to the size of the monetary stakes. Indeed, meta-analytic reviews indicate that game behavior tends to differ rather minimally over variance in stake size (Engel, 2011; Johnson & Mislin, 2011). In addition, both our measure of self-perceived moral superiority, and our analytic approach, were comprehensive—comprising a variety of validated moral traits (see Tappin & McKay, 2017), and a range of robustness checks, respectively. We expect conceptual replications that use alternative measures of moral superiority and alternative analytic approaches to produce similar results to those we observed here. We have no reason to believe that the results depend on other characteristics of the participants, materials, or contexts (Simons et al., 2017).

Here we investigated how self-perceived moral superiority related to moral behavior as measured in canonical economic games. We observed robust evidence that self-perceived

moral superiority is not associated with magnitude of trust in others, trustworthiness, or fairness, as defined by the games; a result seemingly produced by the opposite behavioral manifestations of (i) self-knowledge and (ii) cynicism about the morality of the average person.

## References


Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: what counts? *Philosophical Psychology, 26,* 673-701. Doi: 10.1080/09515089.2012.677397

Alicke, M. D., & Sedikides, C. (Eds.). (2011). *Handbook of self-enhancement and self-protection*. New York, NY: Guilford Press.

Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of $1 stakes. *PloS One, 7,* e31461. Doi: 10.1371/journal.pone.0031461

Arechar, A. A., Gaechter, S., & Molleman, L. (in press). Conducting interactive experiments online. *Experimental Economics*, 1-33. Doi: 10.1007/s10683-017-9527-2

Barranti, M., Carlson, E. N., & Furr, R. M. (2016). Disagreement About Moral Character Is Linked to Interpersonal Costs. *Social Psychological and Personality Science, 7,* 806-817. Doi: http://dx.doi.org/10.1177%2F1948550616662127

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122-142. Doi: 10.1006/game.1995.1027

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field? Evidence from donations. *Experimental Economics, 11,* 268-281. Doi: 10.1007/s10683-007-9192-y

Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology, 76,* 420-430. Doi: 10.1037/0022-3514.76.3.420

Brown, J. D. (2012). Understanding the better than average effect motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209-219. Doi: 10.1177/0146167211432763

Brülhart, M., & Usunier, J. C. (2012). Does the trust game measure trust? *Economics Letters, 115,* 20-23. Doi: 10.1016/j.econlet.2011.11.039

Capraro, V., & Rand, D. G. (2017). Do the Right Thing: Preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Available at SSRN*: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2965067

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130,* 813-838. Doi: 10.1037/0033-2909.130.5.813

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26,* 1131-1139. Doi: http://dx.doi.org/10.1177%2F0956797615585115

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12,* 53-81. Doi: 10.1146/annurev-clinpsy-021815-093623

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics, 14,* 583-610. Doi: 10.1007/s10683-011-9283-7

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction? *Journal of Personality and Social Psychology, 79,* 861-875. Doi: 10.1037/0022-3514.79.6.861

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149-1160. Doi: 10.3758/BRM.41.4.1149

Fehr, E., & Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics, 95,* 1144-1155. Doi: 10.1016/j.jpubeco.2011.05.013

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*, 817-868. Doi: 10.1162/003355399556151

Fetchenhauer, D., & Dunning, D. (2006). Perceptions of prosociality and solidarity in self and others. In D. Fetchenhauer, A. Flache, B. Buunk, & S. Lindenberg (Eds.), *Solidarity and prosocial behavior* (pp. 61-74). New York, NY: Springer.

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology*, *30*, 263-276. Doi: 10.1016/j.joep.2008.04.006

Fetchenhauer, D., & Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science*, *21*, 189-193. Doi: http://dx.doi.org/10.1177%2F0956797609358586

Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics*, *16*, 155-169. Doi: 10.1007/s10683-012-9337-5

Galizzi, M. M., & Navarro-Martinez, D. (2017). On the external validity of social preference games: a systematic lab-field study. *Management Science*.

Heck, P. R., & Krueger, J. I. (2015). Self-enhancement diminished. *Journal of Experimental Psychology: General*, *144*, 1003-1020. Doi: 10.1037/xge0000105

Heck, P. R., & Krueger, J. I. (2016). Social perception of self-enhancement bias and error. *Social Psychology*, *47*, 327-339. Doi: 10.1027/1864-9335/a000287

Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, *11*, 4-27. Doi: http://dx.doi.org/10.1177%2F1088868306294587

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Lesorogol, C. (2006). Costly punishment across human societies. *Science*, *312*, 1767-1770. Doi: 10.1126/science.1127333

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, *54*, 598-603. Doi: 10.1016/j.paid.2012.11.008

JASP Team (2017). JASP (Version 0.8.1.1) [Computer software].

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865-889. Doi: 10.1016/j.joep.2011.05.007

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*. *Available at SSRN*: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2897313

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology*, *110*, 660-674. Doi: 10.1037/pspa0000050

Klein, N., & Epley, N. (in press). Less evil than you: Bounded self-righteousness in character inferences, emotional reactions, and behavioral extremes. *Personality and Social Psychology Bulletin.* Doi: 10.1177/0146167217711918

Krueger, J. I., & Acevedo, M. (2007). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *The American Journal of Psychology*, *120*, 593-618. Doi: http://www.jstor.org/stable/20445427

Krueger, J. I., & DiDonato, T. E. (2010). Perceptions of morality and competence in (non) interdependent games. *Acta Psychologica*, *134*, 85-93. Doi: 10.1016/j.actpsy.2009.12.010

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence to social norms. *Negotiation and Conflict Management Research*, *1*, 31-52. Doi: 10.1111/j.1750-4716.2007.00003.x

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54*, 1053-1090. Doi: 10.1037/0003-066X.54.12.1053

O'Mara, E. M., & Gaertner, L. (2017). Does self-enhancement facilitate task performance? *Journal of Experimental Psychology: General*, *146*, 442-455. Doi: 10.1037/xge0000272

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications, 5:4939*. Doi: 10.1038/ncomms5939

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172-179. Doi: 10.1016/j.jtbi.2011.03.004

Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General*, *145*, 389-396. Doi: 10.1037/xge0000154

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sedikides, C., Gaertner, L., & Cai, H. (2015). On the Panculturality of Self-enhancement and Self-protection Motivation: The Case for the Universality of Self-esteem. In A.J. Elliot (Ed.), *Advances in Motivation Science* (pp. 185–241).

Sedikides, C., Meek, R., Alicke, M. D., & Taylor, S. (2014). Behind bars but above the bar: Prisoners consider themselves more prosocial than non-prisoners. *British Journal of Social Psychology*, *53*, 396-403. Doi: 10.1111/bjso.12060

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology*, *29*, 209-269. Doi: 10.1016/S0065-2601(08)60018-0

Simons, D. J., Shoda, Y., & Lindsay, S. D. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, 1-6. Doi: 10.1177/1745691617708630

Simpson, B., Harrell, A., & Willer, R. (2013). Hidden paths from morality to cooperation: Moral judgments promote trust and trustworthiness. *Social forces*, *91*, 1529-1548. Doi: muse.jhu.edu/article/509346

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159-171. Doi: 10.1016/j.cognition.2013.12.005

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, *26*, 1469-1479. Doi: http://dx.doi.org/10.1177%2F0956797615592381

Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science*. Doi: 10.1177/1948550616673878

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193-210. Doi: 10.1037/0033-2909.103.2.193

Van Damme, C., Deschrijver, E., Van Geert, E., & Hoorens, V. (2017). When Praising Yourself Insults Others: Self-Superiority Claims Provoke Aggression. *Personality and Social Psychology Bulletin*. Doi: 10.1177/0146167217703951

Van Damme, C., Hoorens, V., & Sedikides, C. (2016). Why self-enhancement provokes dislike: The hubris hypothesis and the aversiveness of explicit self-superiority claims. *Self and Identity*, *15*, 173-190. Doi: 10.1080/15298868.2015.1095232

Van Lange, P. A., & Sedikides, C. (1998). Being more honest but not necessarily more intelligent than others: Generality and explanations for the Muhammad Ali

effect. *European Journal of Social Psychology*, *28*, 675-680. Doi: 10.1002/(SICI)1099-0992(199807/08)28:4<675::AID-EJSP883>3.0.CO;2-5

Vazire, S., & Carlson, E. N. (2010). Self-Knowledge of Personality: Do People Know Themselves? *Social and Personality Psychology Compass*, *4*, 605-620. Doi: 10.1111/j.1751-9004.2010.00280.x

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin*, *90*, 245-271. Doi: 10.1037/0033-2909.90.2.245

# Supplementary Information:

## *Investigating the relationship between self-perceived moral superiority and moral behavior using economic games*

## Section 1

### Bayesian analyses

All Bayesian analyses were Kendall's tau Bayesian correlation pairs conducted in JASP. Below we report the robustness checks (displayed graphically) for each Bayesian analysis reported in the main text.

### Self-perceived moral superiority and trust in others

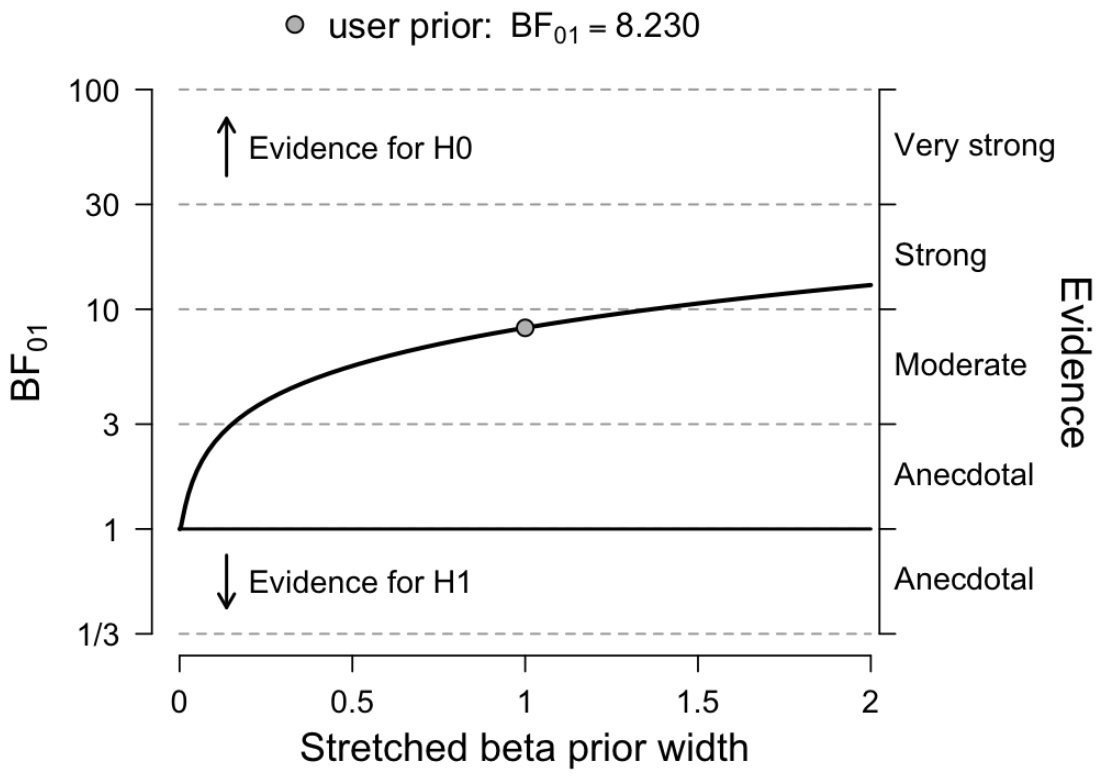


**Figure S1 | Bayes Factor robustness check: Trustors.** Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.

**Self-perceived moral superiority and trustworthiness**

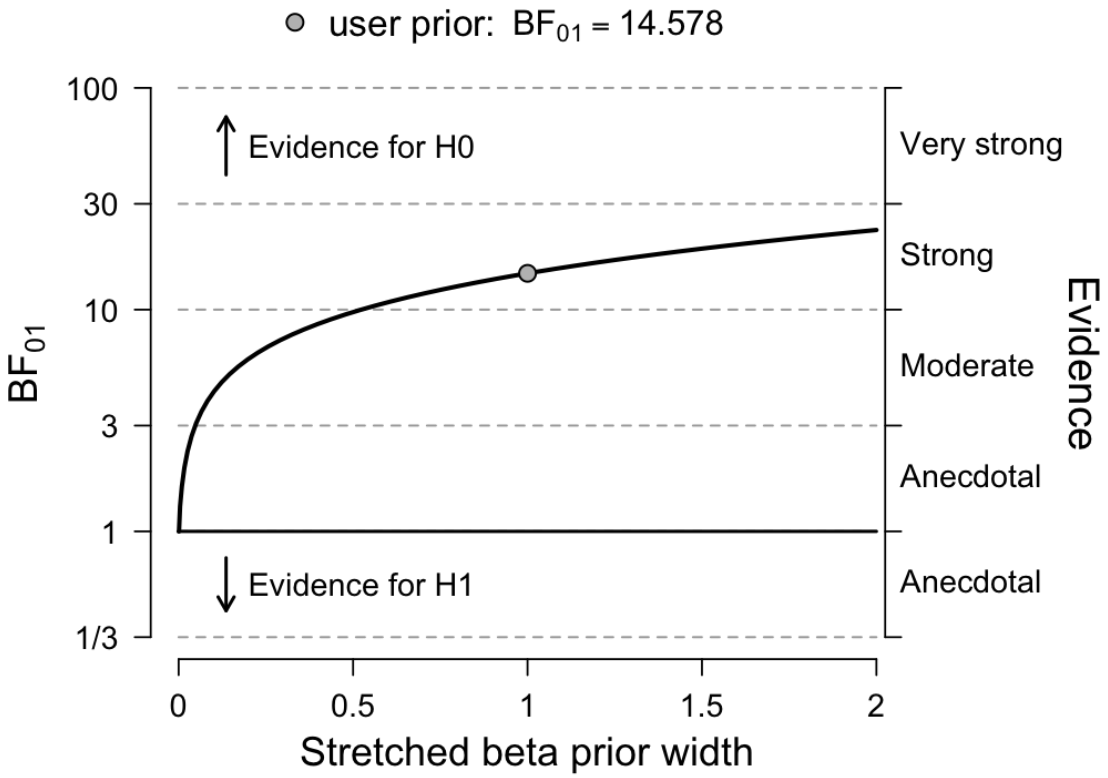


**Figure S2 | Bayes Factor robustness check: Trustees.** Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.

**Self-perceived moral superiority and fairness**

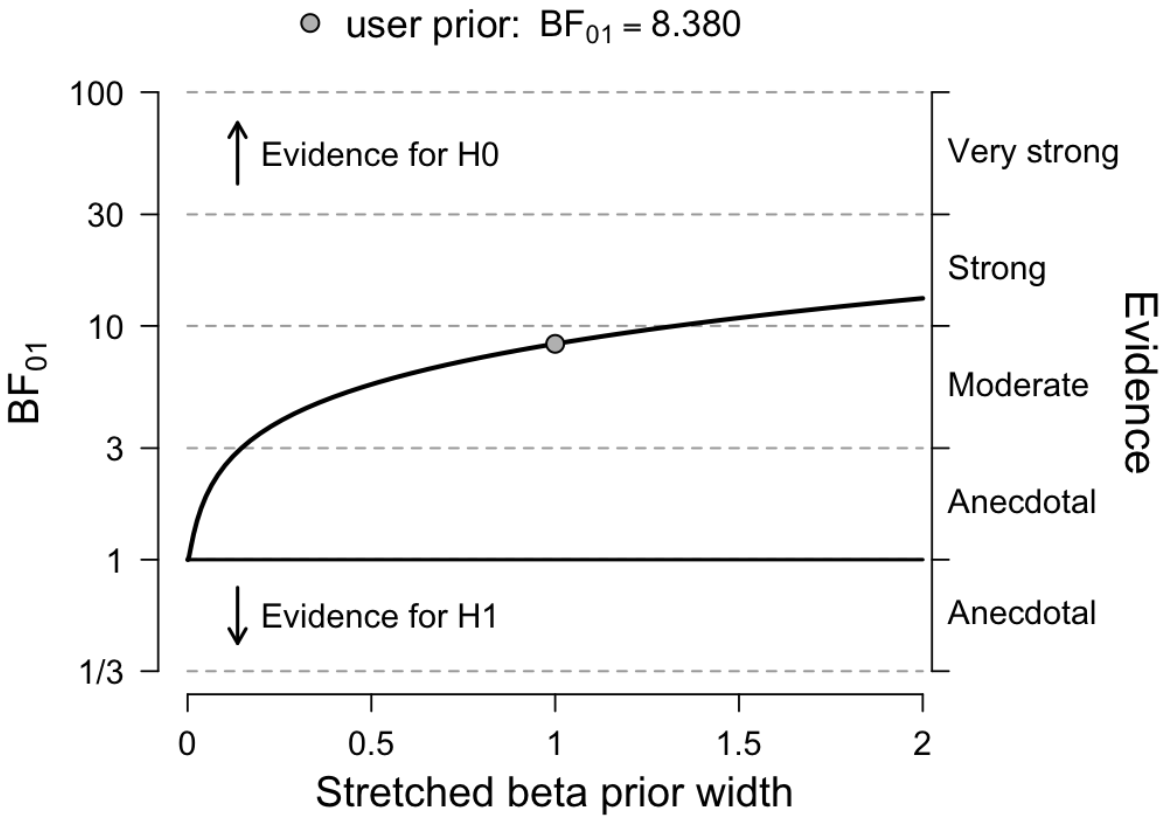


**Figure S3 | Bayes Factor robustness check: Dictators.** Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.

**Self-perceived moral superiority and fairness (truncated sample)**

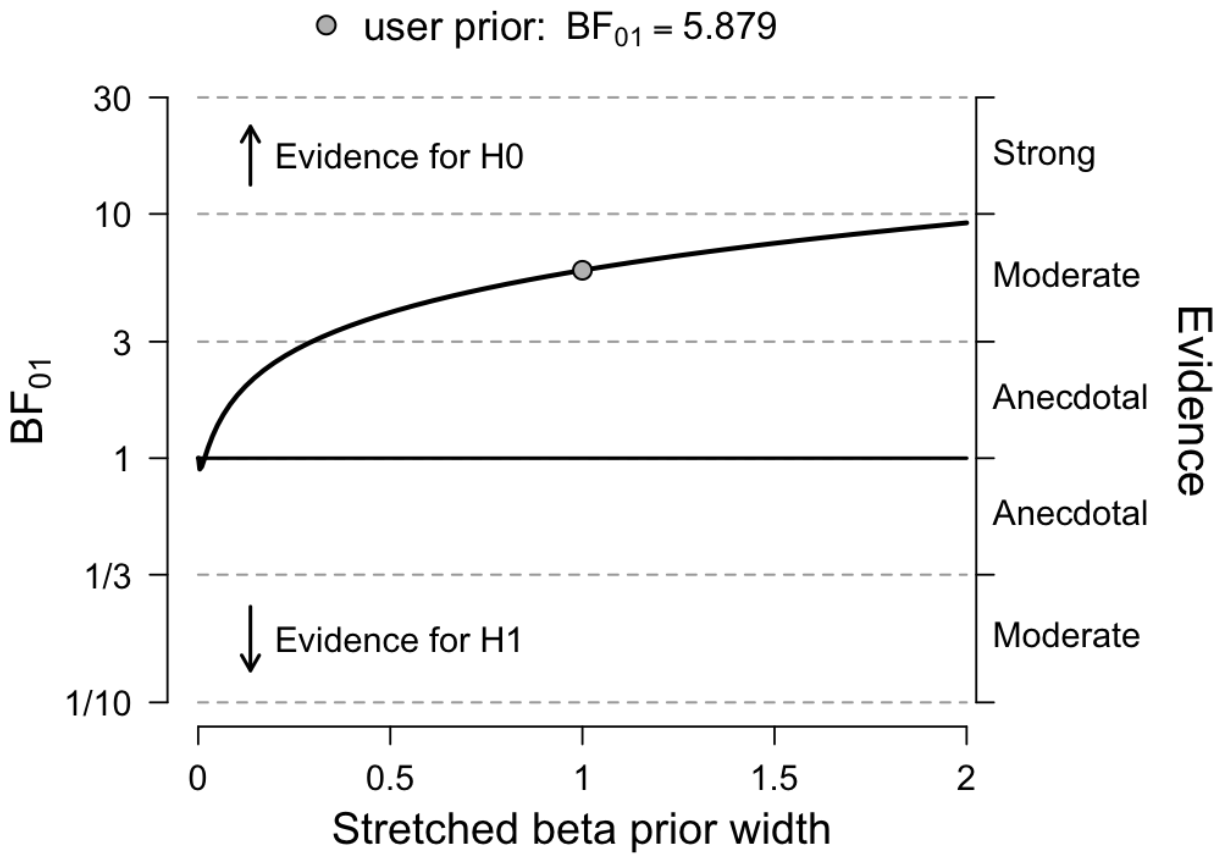


**Figure S4 | Bayes Factor robustness check: Dictators (truncated).** Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.

## Section 2

**Previous experience with economic games**

For all DVs (trustor transfers, trustee transfers, dictator transfers), we explored whether prior experience with the games was masking an association between self-perceived moral superiority and decision behavior (cf. Chandler et al., 2015). Below we present these analyses in the order of DVs present in the main text.

Excluding trustors who reported having previously seen the TG ($N_{excluded}$=125, 33.88%), and repeating the preregistered regression analysis, corroborated the preregistered and other exploratory analyses: F (1, 242) = 0.41, p=.522, R=.04 [b=-0.66, se=1.03, t=-0.64]. Similarly, excluding trustees who reported having previously seen the TG ($N_{excluded}$=114, 31.06%), and repeating the preregistered regression analysis revealed similar results to those in the full (nonnaive) sample: F (1, 251) = 0.01, p=.929, R=.01 [b=0.35, se=3.89, t=0.09]. Finally, excluding dictators who reported having previously seen the DG ($N_{excluded}$=140, 37.94%), and repeating the preregistered regression analysis, produced the same pattern of results as those in the full (nonnaive) sample: F (1, 227) = 0.35, p=.555, R=.04 [b=-0.64, se=1.08, t=-0.59]. All these exploratory analyses are consistent with the preregistered and exploratory analyses reported in the main text; specifically, indicating that prior experience with the economic games was not masking an association between our variables of interest.

## Section 3

### The domain-generality of our measure

To explore the possibility that the domain-generality of the self-perceived moral superiority measure was obscuring any relationship between perceived superiority on *specific* moral traits (e.g., trustworthiness, or fairness) and behavior representative of those traits, we first computed difference scores between participants' given self-judgments (s) and their judgments of the average person (o) for the traits "trustworthy" and "fair" only. We then conducted Spearman's rank correlations between these scores and trustee decisions (Study 1), and dictator decisions (Study 2), respectively. Perceived superiority on the trait "trustworthy" was trivially related to trustee back-transfer amount: $r_s$ = -.03, p = .588. Similarly, perceived superiority on the trait "fair" was trivially related to dictator transfer amount: $r_s$ = -.002, p =

.973. These results mirror those of the preregistered analyses using the self-perceived moral superiority measure.

## Section 4

**Explaining variance in self-perceived moral superiority**

We explored whether self-perceived moral superiority scores were better explained by (i) moral self-perceptions ($b_{SD}$), or (ii) perceptions of the average person's morality ($b_{OD}$). Separately correlating (i) and (ii) with self-perceived moral superiority scores indicated that the latter explained, on average, 64.20% variance in these scores (Study 1: r = -.79, p <.001, 62.16% variance explained, Study 2: r = -.81, p <.001, 66.24% variance explained). Whereas, the former accounted for less than a quarter of this amount (average variance explained: 14.06%, Study 1: r = .35, p <.001, 12.35% variance explained, Study 2: r = .40, p <.001, 15.77% variance explained).

## Section 5

**Defensible self-perceived moral superiority and economic game behavior**

We explored whether the "defensible" component of self-perceived moral superiority—as given by the regression-based index—was associated with behavior in the economic games. This component is defined by $b_{ID} - b_{OD}$; or, the amount of self-superiority that may be justified by the fact that individuals have limited information about the average person (Heck & Krueger, 2015; Tappin & McKay, 2017). Defensible self-perceived moral superiority was weakly but positively correlated with transfer amount for those in the role of trustor [$r_s$ = .11, p=.034], trustee [$r_s$ = .12, p=.023], and dictator [$r_s$ = .08, p=.110]. This provides some intuitive rationale for labeling the index "defensible", but we emphasize that caution must be used when interpreting the meaning of these results given the lack of a priori predictions we had about these associations.

## Section 6

**Replication of Tappin & McKay (2017)**

We included 20 nonmoral filler traits in the trait judgment task—10 of which pertained to the domain of agency, and 10 to the domain of sociability (also drawn from Tappin & McKay, 2017)—and we were thus able to replicate the primary results reported in Tappin and McKay (2017) (Table S1 displays the full list of traits). Specifically, in their study they found that self-perceived moral superiority—measured using the same regression-based index as in the current studies—was larger in magnitude, and more frequent, than perceived superiority in *nonmoral* domains of social perception. Thus, in the following section we reproduce the primary analyses reported in Tappin and McKay (2017) (p.6, para beginning: "*This indicates that irrational self-enhancement is strongest in the moral domain.*")[4]. For consistency, we use their terminology to refer to perceived superiority (that is, "*irrational self-enhancement*").

[4] Note: we replicate the *magnitude* and *frequency* analyses of Tappin and McKay (2017) only, not the analyses with self-esteem (since we did not collect self-esteem data in the current studies). Also, sample N's differ from those reported in the preregistered analyses because, prior to replicating Tappin & McKay (2017), we had to additionally exclude uniform responders in the nonmoral trait domains.

Table S1.

Full list of traits used in Studies 1 and 2

| Domain | Positive traits | Negative traits |
|---|---|---|
| Morality | Honest | Insincere |
| | Trustworthy | Prejudiced |
| | Fair | Disloyal |
| | Respectful | Manipulative |
| | Principled | Deceptive |
| Agency | Hard-working | Lazy |
| | Knowledgeable | Undedicated |
| | Competent | Unintelligent |
| | Creative | Unmotivated |
| | Determined | Illogical |
| Sociability | Sociable | Cold |
| | Playful | Disagreeable |
| | Warm | Rude |
| | Family-orientated | Humorless |
| | Easy-going | Uptight |

*Note.* Traits were identical to those used in Tappin & McKay (2017), except that the trait "playful" replaced "cooperative".

***Irrational self-enhancement.*** We first investigated the magnitude of irrational self-enhancement in each trait domain by examining how well trait desirability predicted actual self-judgments (mean $b_{SD}$) vs. inferred self-judgments (mean $b_{ID}$). In both studies, replicating Tappin and McKay (2017), paired t-tests revealed that magnitude of irrational self-enhancement was largest in the moral domain, *Study 1*: **Morality** (0.74 vs. 0.23), t (727) = 33.52, p <.001, Cohens **d = 1.24** 95% Confidence Interval [1.15, 1.34], **Agency** (0.73 vs. 0.28), t (727) = 24.50, p <.001, **d = 0.91** [0.82, 0.99], and **Sociability** (0.61 vs. 0.52), t (727) = 4.17, p <.001, **d = 0.15** [0.08, 0.23]; *Study 2*: **Morality** (0.76 vs. 0.21), t (361) = 23.60, p <.001, **d = 1.24** [1.10, 1.38], **Agency** (0.72 vs. 0.28), t (361) = 17.10, p <.001, **d = 0.90** [0.78, 1.02], and **Sociability** (0.63 vs. 0.49), t (361) = 5.49, p <.001, **d = 0.29** [0.18, 0.39].

We confirmed statistically that Morality was the largest by computing the difference measure ($b_{SD} - b_{ID}$) for each trait domain, and conducting paired t-tests *between* trait domains. In both studies, replicating Tappin and McKay (2017), the moral domain comprised the largest magnitude of irrational self-enhancement, *Study 1*: **Morality** (0.52) vs. **Agency** (0.46), t (727) = 4.00, p <.001, **d = 0.15** [0.08, 0.22], and vs. **Sociability** (0.09), t (727) = 25.36, p <.001, **d = 0.94** [0.85, 1.03]; *Study 2*: **Morality** (0.55) vs. **Agency** (0.44), t (361) = 5.04, p <.001, **d = 0.27** [0.16, 0.37], and vs. **Sociability** (0.15), t (361) = 17.16, p <.001, **d = 0.90** [0.78, 1.02].

Finally, corroborating the analysis of magnitude, and again replicating Tappin and McKay (2017), McNemar's Tests showed that more individuals irrationally self-enhanced ($b_{SD} > b_{ID}$) in the moral domain than in either of the nonmoral domains, *Study 1*: **Morality** (n = 659, 90.52%) **vs. Agency** (n = 611, 83.93%), $\chi^2$ (df = 1, N = 728) = 26.94, p <.001, and **vs. Sociability** (n = 396, 54.40%), $\chi^2$ (df = 1, N = 728) = 249.61, p <.001; *Study 2*: **Morality** (n = 331, 91.44%) **vs. Agency** (n = 302, 83.43%), $\chi^2$ (df = 1, N = 362) = 16.68, p <.001, and **vs. Sociability** (n = 214, 59.12%), $\chi^2$ (df = 1, N = 362) = 109.40, p <.001.