SOCIAL OBSERVATION AND MORAL HYPOCRISY

By

Andrew Marcel Defever

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology - Master of Arts

2016

ABSTRACT

SOCIAL OBSERVATION AND MORAL HYPOCRISY

By

Andrew Marcel Defever

Previous research shows that self-interest has a motivating influence in moral reasoning and decision-making. However, cues of social observation show a positive influential effect on moral and cooperative behavior, in both experimental and real-world contexts. Implementing an expected utility framework to model moral deliberation as a cost-benefit calculation, the present study synthesizes these two effects by examining whether social observation impacts decisions in a moral dilemma situation. Utilizing Batson et al.'s (1997) moral dilemma paradigm, we test whether the perceived presence of observers increases the likelihood of making a fair allocation of a reward in a large university sample ($N = 161$). Across three social observation conditions, participants' allocation decisions were recorded, including their emotional reactions and open-ended justifications. Behavioral and affective response patterns indicated that participants acted in accordance with a self-interested, morally hypocritical motivational approach, while cues of observation were not shown to influence behavior. Past and future theoretical implications are discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

The presence of moral behavior is ubiquitous across human cultures, yet discrepancies exist between individuals' morals and their actions. In 2009, a U.S. state representative who fought against the Planned Parenthood initiative because he claimed it promoted premarital sex was caught having an affair with a 22-year-old intern. In 2014, a politician who campaigned against marriage equality was discovered performing as a drag queen at a local bar. Colloquially described as not 'practicing what you preach', moral hypocrisy is the inconsistency between an individual's moral values and their actions.

Moral hypocrisy is a complex, multi-faceted phenomenon involving a mixture of behavior, outcomes, and intentions. We focus here on the paradigmatic case of hypocrisy defined as behavioral inconsistency, or claiming a moral position but acting in contradiction to it (Monin & Merritt, 2011). Such hypocrisy has been demonstrated empirically, most notably by Daniel Batson and colleagues across a series of studies examining individuals' choices in a real-time moral dilemma. Participants were asked to assign a rewarding task and a neutral task amongst themselves and another person. Of those who opted for a 'fair' approach and flipped a coin to assign the tasks, most (85-90%) selfishly assigned the rewarding task to themselves. This is contrary to the 50% assignment rate expected if individuals were actually using the coin as a fair method. Further, participants who flipped the coin rated themselves as more proud, more moral, and less guilty about their task assignment decision than those who did not flip the coin and simply assigned the rewarding task to themselves. In other words, they gave the appearance of moral intentions by flipping the coin, while subsequently acting in a self-interested manner (Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997; Batson, Thompson, & Chen, 2002; Batson, Thompson, Seuferling, Whitney, & Strongman, 1999).

Why might a person behave hypocritically? Plausibly, there must be something to be gained from the decision, as there are costs associated with violating moral rules and customs. When someone is caught behaving unfairly, for instance, it can produce feelings of contempt, disgust, and anger in others (Haidt, 2003). Further, second and third parties have been shown to initiate punishment behavior against uncooperative or unfair exchange partners, with evidence showing that negative affective responses to non-cooperators mediate the relationship between violation and punishment (Fehr & Gächter, 2002; Fehr & Fischbacher, 2004). Under the threat of punishment, then, hypocrisy allows a person to endorse a particular course of action to preserve their moral reputation and avoid the costs of punishment, while continuing to pursue the benefits associated with the violation.

The phenomenon of hypocrisy is well supported, both observationally and empirically. However, not everyone acts in a hypocritical manner, and individuals often act in congruence with their morals. This begs the question of what conditions or contextually relevant factors may facilitate a hypocritical course of action? The present study seeks to examine whether the perceived threat of sanctions influences the probability of engaging in a morally congruent versus a self-interested or hypocritical course of action. Specifically, utilizing Batson et al.'s (1997, 1999) experimental paradigm, we examine whether the perception of being observed by others - posited as an index of the likelihood of being punished for a moral violation - influences the likelihood of acting fairly in a situation that pits moral congruency against self-interested gains. We implement expected utility theory as a predictive framework, modeling the decision to act in an unfair manner for selfish gain as a cost-benefit calculation, mediated by the probability of being observed and punished for such behavior.

Expected Utility and Morality

The outcome tradeoffs inherent to the moral decision process can be modeled using an economic cost-benefit framework. Decision theory describes how people decide amongst alternative choices, based on the expected utility (EU) of the outcome for each choice (Fishburn, 1982). Individuals pursue the action with the highest utility, weighted by the probability of each outcome occurring. When conceptualizing the costs and benefits for a morally relevant decision, punishment is a probabilistic calculation, the likelihood of which is not necessarily 100%. It is contingent upon at least one person obtaining incriminating evidence and subsequently pursuing a punitive course of action. This creates a situation of uncertainty where agents must estimate the likelihood of success or failure for a particular courses of action, taking into account relevant socioecological factors at the moment of action. For example, whether others are present in the immediate environment, the conspicuousness of the act, and whether they can observe your transgression. Subsequently, when estimating the relative value of action outcomes in an EU framework, the uncertainty of punishment manifests as an attenuation of the estimated costs of a moral transgression.

When modeling this deliberative process, individuals can either take action or not take action, and they are either caught for the moral transgression or they get away with it. The uncertainty of being caught qualifies the utility value of each outcome as a probability value ranging from 0.0 (no perceived possibility of being caught) to 1.0 (complete certainty of being caught). The punishment modifier provides a dynamic component which allows for a more accurate estimate of the relevant outcome tradeoffs. For example, consider someone who is contemplating stealing from one of two different corner stores. The first store has visible security cameras mounted around the store, while the other does not. Assuming all other things equal, the

rational decision would be to steal from the second store with no cameras. Even though in both situations the certainty of being caught would not necessarily be 100%, the likelihood of detection would be perceived as higher if there were security cameras present.

Figure 1 outlines the variables in the EU model: the benefit (*b*) of a successful action, the probability (*p*) of being caught, and the cost (*c*) for being punished when caught. The utility for no action is inherently zero. There is nothing to be gained, and no risk of punishment, if no immoral act is committed in the first place. The utility for action is the sum of the values for being caught and not being caught:

$$EU(action) = b(1-p) + c(p)$$

If EU(action) is greater than zero, the deliberator should take action, because the value of taking action is (probabilistically) higher than inaction. This certainly does not guarantee success for any particular decision, but over time individuals who consistently choose actions with higher utility should, on average, fare better than those who do not.

From this model we can derive basic predictions about how individuals evaluate a situation and ultimately decide to take action. Increasing *b*, or decreasing *c* and/or *p,* should increase EU(action) and subsequently the likelihood of action. For example, the prospect of a stolen $1000 is worth more than a stolen $50, so we would expect more people to take action when the potential benefit is higher. A fine of $200 is less costly than 5 years in prison, so we would expect less people to take action when potential costs are increased.

The relative values of *b* and *c* are subjective within any one person, creating variance across individuals in the same situation. For example, a stolen $50 might mean more to someone who is unemployed compared to a millionaire, or the reputational costs for lying might mean more to a public figurehead than to an average citizen. Further, a person's circumstances at any

given time could influence their relative likelihood of action. This makes it difficult to predict what a single person might do in a given situation. However, we would expect to see predictable mean changes across groups of individuals in their propensity for action, based on the parameter changes outlined above.

Social Tracking. The value of *p* is also dynamic; a direct index of the probability of each outcome occurring. We theorize that with morally relevant decisions, *p* is dependent on the perception of environmental cues indicative of being observed during the transgressive act. Humans are competent social detectives. They can discern cheating behavior during social exchanges, and track the intentions and motivations of others' actions (Cosmides & Tooby, 2005; Cosmides, Tooby, Fiddick, & Bryant, 2005). For example, laboratory studies have demonstrated that when individuals make judgments in moral dilemmas, they rate intentional harm as morally worse than harm through omission (DeScioli, Asao, & Kurzban, 2012; DeScioli, Christner, & Kurzban, 2011), and find harm used as a means to an end morally worse than harm that occurs as a by-product of action (Greene et al., 2009; Mikhail, 2007). Further, developmental evidence has shown that infants and babies prefer individuals who demonstrate helping behavior towards others (Bloom, 2013), and children as young as 4 years old have demonstrated intent-based moral judgment of others (Cushman, Sheketoff, Wharton, & Carey, 2013).

Consequently, humans are keen trackers of cues related to social privacy and observation. Of particular significance, individuals are highly sensitive to watching eyes, a direct indicator of the presence of another person in the immediate environment. The presence of eyes has been shown to increase cooperative behavior in economic games (Haley & Fessler, 2005) and reduce antisocial behavior such as littering and theft (Ernest-Jones et al., 2011; Nettle, Nott, & Bateson,

2012). This includes starkly non-human entities with eyes (Burnham & Hare, 2007), and even minimal cues consisting simply of three dots in a 'watching eyes' face configuration (Rigdon et al., 2009). Within the context of the EU framework, social observer tracking allows individuals to more accurately assess the potential risk ($p*c$) inherent to contexts of uncertainty, where the presence of others may lead to punishment for a particular choice of action.

Evolution of Social Tracking. Evolutionary theory posits that sensitivity to social cues may have evolved as an adaptive response to the threat of punishment from others. Ethnographically, humans have a long history of interpersonal cooperation. However, cooperative groups have the potential to facilitate free riding behavior, defined as the consumption or use of a publically-derived good without contributing (or contributing less) to its production. To solve this dilemma, humans are thought to have evolved a propensity to sanction non-cooperators and norm violators within their group (Boehm, 2012; Boyd, Gintis, Bowles, & Richerson, 2003; Henrich et al., 2006), in the form of physical sanctions (e.g., harm or death), social sanctions (reputational damage, ostracism), or a combination of the two, implemented by second party victims (retaliation) or third party observers (altruistic punishment; DeScioli & Kurzban, 2009).

Punishment is thought to have imparted consistent adaptive pressure for cooperative behavior in early humans and contemporary hunter-gatherer groups (Boehm, 2012). Such consistent pressure can induce the evolution of behavioral adaptations (Boyd & Richerson, 1992; Haselton & Buss, 2000; Johnson, Blumstein, Fowler, & Haselton, 2013), such as a propensity to punish moral violations or cognitive mechanisms designed to track others in the environment who may impart such punishment, as a way to reduce fitness costs associated with certain behaviors. Consistent with this line of reasoning, individuals in cooperative contexts show a

propensity to punish moral norm violations, and their tendency toward cooperative behavior show a proximate sensitivity to situations where punishment can and cannot be imparted. For instance, when the classic dictator game (DG) is designed to allow participants to punish others for not offering a fair portion of an initial resource endowment, observers uninvolved in the transaction are willing to spend some of their own resources to punish them (Fehr & Fischbacher, 2004). Further, the presence of the ability to punish others in a public goods game (PGG) substantially increased cooperation amongst individuals in small groups (via significantly larger contributions to a 'public good' resource pool), while removing the ability to punish substantially decreased cooperation (Fehr & Gächter, 2002).

The Present Research

Moral decision-making is a cost-benefit tradeoff with uncertain outcomes, qualified by the probability of actually incurring the costs related to a transgression via social observation and subsequent sanctions. As such, moral hypocrisy is facilitated through the dilemma of desiring the most valuable outcome while attempting to avoid punishment for a moral transgression. Batson and colleagues demonstrated the phenomenon of hypocrisy empirically through a series of experiments pitting selfish gains against a moral sense of fairness. Here we seek to investigate whether cues of social observation attenuate the likelihood of participants acting in a selfish manner, by adding controlled cues of social observation serving as an index of *p*. Implementing the EU framework within this allocation dilemma, we predict that as perceived observation increases, the likelihood of acting in a fair and morally consistent manner should increase in proportion to the perceived likelihood of detection.

Manipulating detection. The most salient indicator of detection is the visual presence of other people. By manipulating the number of perceived observers present during the task

assignment decision, the likelihood of detection (*p*) should increase relative to the number of observers present. Here we utilize mirrors to induce the perception of being observed by others. In their 1999 study, Batson and colleagues had participants engage in the task allocation dilemma in front of a mirror. This was the only condition across all their studies where hypocrisy was not present in their findings. Participants who flipped the coin assigned the rewarding task to themselves at a rate of 50% (as would be expected by a fair coin flip). They interpreted the results as the mirror increasing participants' self-awareness, serving to highlight the discrepancy between participants' actions and their internalized sense of fairness and bringing their behavior in line with their morals (Wicklund, 1975). However, an evolutionary explanation suggests that individuals were not simply seeing themselves in the mirror, but perceived another observer watching their behavior.

As outlined above, in the environment of evolutionary adaptedness the threat of punishment from others is posited to have facilitated adaptations designed to track cues of observation and detection in the environment. Early humans were not exposed to mirrors or reflective surfaces as consistently as we are today, and likely would not have evolved a specific cognitive mechanism sensitive to the input of one's own reflection. Indeed, the regular viewing of one's own image is probably a relatively novel and recent phenomenon in human history, as the first manufactured mirror is speculated to have been constructed as recently as 8,000 years ago (Enoch, 2006).

From a functional perspective, in Batson et al.'s (1999) study when participants saw a human figure or, more specifically, a pair of watching eyes in the mirror during the experiment, this may have activated the social observation detection system, tracking the presence of watching eyes during their behavior. This would have created a perception of being observed,

and according to the EU model individuals would have perceived the costs of selfish behavior to be higher due to the increase in *p*. Subsequently, less individuals would be expected to act selfishly and unfairly. Mirrors have been shown to influence behavior in other morally relevant domains as well, most notably in the reduction of cheating behavior (Diener & Wallbom, 1976; Heine, Takemoto, Moskalenko, Lasaleta & Henrich, 2008). Thus, we propose that using one or more mirrors will serve to induce the perception of being observed in the experiment, simulating a situation where participants will feel as though they are being watched by one or more persons while they are performing the experimental procedures.

Moral Emotions. In order to gauge how people are reacting to our moral dilemma on an individual level, we implement positive and negative emotion measurements to gauge the affective responses of participants in the study. Tangney, Stuewig, and Mashek (2007) describe guilt, shame, and pride as three primary moral emotions. Guilt and shame are negatively valenced, primarily evoked by morally relevant transgressions. Guilt experiences focus on how a specific action impacts others, while shame experiences focus on how others evaluate the self. Pride, a positively valenced moral emotion, is defined as an emotion evoked by appraisal of one's actions as socially valuable or that one is a socially valuable individual (Mascolo & Fischer, 1995).

In a morally salient situation, then, moral emotions can serve as an 'affective barometer' – a feedback system reflecting the internal responses of a moral agent in a particular situation. Emotions are thought to stem from an evolutionarily ancient system, with humans showing innate and universal emotional expressions cross-culturally (Ekman, 1993; 2007). Limited evidence shows that observation can potentially interfere with emotional processing systems in the brain (Yu, Muggleton & Juan, 2015), or inhibit externalizations when others are present (for

example, see Friedman & Miller-Herringer, 1991). However, based on the limited evidence, in the present study we remain agnostic as to the influence social observation may have on emotion.

Predictions

In the experiment, participants are faced with the dilemma of using a fair method (a coin flip) to allocate a rewarding task between themselves and another person. Previous investigations found that individuals who flipped the coin were doing so to *appear* moral, as the majority of participants (greater than 50%, as would be expected by chance) who flipped still assigned themselves the rewarding task (Batson et al., 1997; Batson et al., 2002; Batson et al., 1999).

Thus, flipping the coin, regardless of the task assignment, is a display of moral intent. Participants give the impression that they are motivated to act fairly, even if they do not follow through with using the coin properly. The EU model predicts that when cues of observation are present, participants should be less likely to pursue selfish gains due to the threat of being observed acting unfairly. Thus, regarding the first decision whether to flip the coin or not, prediction (1) states:

(1) *Participants will be significantly more likely to flip the coin in the low and high observation conditions relative to the no observation condition, and will be most likely to flip in the high observation condition.*

For those who choose not to flip the coin, rates of assignment of the participant to the positive task are anticipated to be fairly high, reflecting participants' desire for the positive task and lack of concern for using a fair method (Batson et al., 1997; Batson et al., 2002; Batson et al., 1999).

The next decision participants face is how to assign the tasks. Cues of observation are posited to lead to the perception of a higher likelihood of punishment. The EU model predicts

that as the likelihood of observation increases, participants will be less likely to assign themselves to the positive task. Thus, regarding the second decision of how to assign the tasks, prediction (2) states:

(2) *Participants will be significantly less likely to assign themselves the positive task in the low and high observation conditions, relative to the no observation condition, and participants will be least likely to assign themselves the positive task in the high observation condition.*

If there is a significant threat of being observed and punished, individuals should subsequently be more likely to use the coin fairly as observation increases, due to an increase in the EU estimate of cheating the coin flip relative to acting fairly. Thus, prediction (3) states:

(3) For those who flip, assignment of the participant to the positive task should more closely approximate 50% (what would be expected by chance outcome from a fair coin flip) in the low and high observation conditions, relative to the no observation condition.

On an individual level, emotional reactivity should reflect how a person feels regarding their decision in the moral dilemma situation. Consistent with previous findings (Batson et al., 1997) regarding the affective reactions of participants to the task assignment decision prediction (4) states:

(4) *Participants who flip the coin relative to those who do not, as well as those who assign the other person to the positive task relative to those who assign themselves to the positive task, will rate themselves as significantly more moral, more proud, less guilty, and less ashamed of their task assignment decision.*

We will also examine how emotional reactivity may change across varying levels of social observation. These will be conducted on an exploratory basis where we will examine both

the main effect of observation and interactions between observation and flip, as well as observation and task assignment.

Methods

Participants

One hundred and ninety-eight undergraduate students volunteered to participate in a study titled "Games, Tasks, and Attitudes" through the online recruitment website at Michigan State University. Students received course credit for their participation. Ages ranged from 18 – 33 (M = 19.7, SD = 1.8), with approximately 69% of participants identifying as female. The racial composition of the sample included 55% White, 24% Asian/Pacific Islander, 13% Black, 4% Hispanic/Latino, 3% Middle Eastern, and 1% other or not specified. Participants were excluded from the analysis for failing the manipulation check (n = 13), technical difficulties involving measurement equipment in the laboratory (n = 5), or not completing all primary outcome measures, including demographic covariates of interest (n = 19). The total sample size included in the analysis was (N = 161).

Materials & Procedure

In consultation with Daniel Batson, identical materials to the 1997 and 1999 moral hypocrisy experiment will be used. This includes all measures, documents, and questionnaires, with original wording and formatting. A summary is provided below, but a full description can be found in Batson et al. (1997; 1999, Study 3; 2002).

Participants arrived at the laboratory for their scheduled time, where they were greeted by a research assistant (RA). Upon entering the lab, they were escorted into a sound dampening isolation booth where they completed all study procedures. Informed consent was provided and completed. Participants then reviewed an introduction form outlining the study procedures: two people (the participant and another person - actually fictitious) were to take part in a study looking at how decision characteristics impact an individual's feelings and reactions after

performing a task, where one of them would be randomly assigned to allocate one task to each of them, a positive (rewarding) task and a neutral task. The participant was always the allocator, and they were informed they would never meet other participant face to face. The positive task involved a monetary incentive in the form of a raffle, where the participant could win tickets towards a drawing for a $30 gift card. The neutral task did not involve this incentive, and was described as generally 'dull and boring'. This created a clear asymmetrical division of the tasks, incentivizing individual rewards on the part of the participant.

Participants were then given a packet of instructions, informing them that they had been randomly selected as the person who was to assign the tasks. A slip of paper was provided where they could indicate which of the two participants was to be assigned to the positive task and the neutral task.

To ensure salience of moral fairness during the task assignment procedure, prior to assigning the tasks participants read a statement indicating that most people consider some sort of even-handed method like flipping a coin the 'fairest way to assign the tasks'. This was included to raise awareness of a moral sense of fairness. However, it was emphasized that participants could still assign the tasks however they desired. A color-coded and labeled coin was included in the packet given to participants, with one side indicating "Self to Positive" and the other side indicating "Other to Positive" (viz. Batson et al., 1999, Study 1).

Following the task assignment procedure, participants completed a series of brief follow-up questionnaires. Measurements of participant's moral emotions, including guilt, pride, and shame, were collected using 7-point Likert scale items, labeled 1 "Not at all", to a midpoint of 4 "Moderately", up to 7 "Extremely". Further, an assortment of 24 additional affective responses were recorded, including happy, anxious, sympathetic, lucky, concerned, softhearted, warm,

distressed, compassionate, upset, tender, moved, worried, disturbed, perturbed, uneasy, relieved, irritated, sad, pleased, afraid, satisfied, unsettled, and calm. These served as both auxiliary individual response measurements, as well as filler items in the questionnaire to mask the salience of the specifically moral emotion items.

To assess how moral participants felt their decision was, they responded to the question, "Do you think the way you made the task assignment was morally right?" on a 9-point Likert scale, ranging from 1 indicating "Not at all" up to 9 indicating "Yes, totally". Open-ended questions assessed what participants thought the most moral way to assign the task was. Finally, demographic characteristics were assessed including age, gender, race, and political orientation. Political orientation was assessed on a 7-point Likert scale, ranging from 1 "Very liberal", to 4 "Neutral, up to 7 "Very conservative". Completion of study procedures took approximately 25-30 minutes. All participants were verbally debriefed after their participation, as well as provided a written debriefing form and a raffle ticket for the drawing (regardless of their task assignment decision).

A video monitoring system was used to record the behavior of participants inside the sound isolation booth. This allowed us to code whether participants flipped the coin, what the result of the coin flip was, and whether they performed any other relevant actions (for example, flipping the coin multiple times). Although the labels on the coin were not clearly visible to the recording equipment, the color coding on the coin allowed for visibility and coding of the coin flip results.

Conditions. Three conditions were implemented: High observation, low observation, and no observation. For the high observation condition, a surrounding array of mirrors was placed inside the booth, one large mirror directly in front and an array of 12 tall 'door' mirrors

surrounding the participant. For the low observation condition, the single large mirror was displayed while the surrounding mirrors were turned around so that the reflective surface was facing away from the participant. In the no observation condition, all of the mirrors were turned around to face away from the participant. In accordance with the original procedures, on the single mirror a sign was located on the bottom left corner which stated, “Mirror for Anderson study, please do not touch”. This helped attenuate suspicion about the conspicuous mirrors in the booth (Batson et al., 1999). Condition was randomly assigned based on repeating sequential order (1-2-3-1-2-3…), such that the next participant that arrived received the next condition on the list.

Results

Table 1 displays the frequency of statements for what participants considered to be the most morally right way to assign the tasks. Approximately 72% of participants stated that the most moral approach would be to either flip the coin (56%) or simply assign the other person to the positive task (16%). Comparing this to the rate of assignment of the self to the positive task (~75% across all conditions; See Table 2), there was a discrepancy between what participants considered to be the most moral approach to the task assignment decision and how they actually assigned the tasks. A total of 63 participants (39%) assigned themselves to the positive task without flipping the coin. However, only 2 participants stated that assigning themselves the positive task was the most morally right approach.

Flipping the coin

Table 2 displays the frequency of participants who flipped the coin, relative to their condition. In the no observation condition, 51% of participants decided to flip the coin. In the low observation condition (single mirror), 65% of participants flipped the coin. In the high observation condition (multiple mirrors), 36% of participants flipped the coin. A Chi-square test indicated a significant difference in the likelihood of flipping the coin across the three conditions, $\chi^2 = 8.52$, $p = .014$.

To directly test whether the likelihood of flipping the coin was attenuated in the single and multiple mirror conditions, relative to the no mirror condition (*prediction 1*), a hierarchical logistic regression was implemented. Coin flip served as the criterion, coded (0 = no flip, 1 = flip). Starting with an intercept-only model (Step 0), we first added condition as a categorical predictor (Step 1), dummy coded with 'no observation' as the reference category. An analysis of the residual deviance indicated the Step 1 model showed a significantly better fit against the null

model, $\Delta D = 8.63$, $\chi^2 = 8.63$, $p = .013$, AIC = 220.55, $R^2_p = .07$. However, there was no significant decrease in the likelihood of flipping the coin in the single or multiple mirror conditions relative to the no observation condition.

To examine whether there was any relationship between demographic characteristics and the likelihood of flipping the coin (Step 2), we entered age ($M = 19.67$, $SD = 1.82$), political orientation ($M = 3.69$, $SD = 1.28$), and gender (69% female) into the model. Gender was dummy coded, with 'male' used as the reference category. Table 3 shows the regression coefficients for the full Step 2 model, including odds ratios (OR) for coefficient estimates and a 95% confidence interval around the estimates. The addition of the demographic variables did not significantly improve model fit over the Step 1 model, $\Delta D = 0.79$, $\chi^2 = 9.42$, *n.s.*, AIC = 225.76, $R^2_p = .08$. Further, there was no significant impact of demographic characteristics on the likelihood of flipping the coin.

Assigning the tasks

Table 2 displays the frequency of participants who assigned themselves to the positive task, relative to condition. In the no observation condition, 75% of participants decided to assign themselves the positive task. In the low observation condition, 71% of participants assigned themselves to the positive task. In the high observation condition, 78% of participants assigned themselves to the positive condition. A Chi-square test indicated no significant difference in the proportion of those assigning themselves to the positive task relative to condition, $\chi^2 < 1$, *n.s*. We also examined whether there was a difference in assignment of the self to the positive task relative to whether participants flipped the coin or not. Again a chi-square test indicated no significant difference in task assignment, $\chi^2 = 1.08$, *n.s.*

To directly test whether the likelihood of assigning the participant to the positive task was attenuated in the single and multiple mirror conditions, relative to the no mirror condition (*prediction 2*), a hierarchical logistic regression was implemented. Assignment of the participant to the positive condition served as the criterion (coded 0 = other to positive, 1 = self to positive). Starting with an intercept-only model (Step 0), we separately added condition (Step 1a) and coin flip (Step 1b) independently as categorical predictors, dummy coded with 'no observation' and 'no flip' as the reference categories, respectively. Step 2 added both variables together, and finally Step 3 added demographic covariates, including age, gender, and political orientation. Gender was dummy coded, with 'male' serving as the reference categories.

An analysis of the residual deviance indicated the Step 3 model showed the best model fit over the null model. However, measures of model fit showed some signs of inadequacy for the current model, $\Delta D = 9.17$, $\chi^2 = 9.18$, $p = .164$, AIC = 187.53, $R^2_p = .08$. Table 4 displays the coefficient estimates, including OR and a 95% confidence interval around the estimates. There was no significant effect of condition or coin flip on the likelihood of assigning the self to the positive condition. There was a significant effect of gender, with females being more likely to assign themselves to the positive task, ($OR = 2.56$, $p = .015$).

Plausible alternatives. We were interested in how observation condition would influence participants' likelihood of assigning themselves to the positive task. However, it could be argued that those who won the coin flip, and subsequently assigned themselves the positive task as a result, added unwanted noise to our analysis by inflating the rate of assignment to the positive condition because they were justified in their assignment of themselves to the positive task. However, upon removing coin flip winners from the analysis, we still did not find a significant effect of observation condition on the likelihood of assigning the tasks. Further, we explored

whether combining the mirror conditions into a single group would result in any change in outcomes. Here we also found no significant effect of observation condition on the outcome of both the coin flip and task assignment. Finally, we explored all interaction effects between predictors of interest and demographic covariates. However, these were not found to be significant, and were subsequently excluded from the models outlined above.

Fair use of the coin

To examine whether the use of the coin was more fair (i.e. more closely approximated the 50% assignment rate expected by chance upon a fair coin flip) in the high and low observation conditions, relative to the no observation condition (*prediction 3*), we examined the rate of assignment of the self to the positive task by condition after sub-setting the data to include only those who flipped the coin ($n = 81$) and examined the frequencies of participants assigning themselves to the positive task. Table 2 shows similar rates of assignment of the participant to the positive task across the no observation ($n = 19$, 68%), low observation ($n = 24$, 73%), and high observation conditions ($n = 14$, 70%). A chi-square test revealed that the difference in assignment rates was not significant across conditions, $\chi^2 = 1.08$, *n.s.*, with an overall rate of assignment of the participant to the positive task ~70%.

Affective responses

We measured a total of 27 affective responses, including happy, anxious, sympathetic, lucky, concerned, softhearted, warm, distressed, compassionate, upset, tender, moved, worried, disturbed, perturbed, uneasy, relieved, irritated, sad, pleased, guilty, proud, afraid, ashamed, satisfied, unsettled, and calm. We dropped lucky from subsequent analysis because it did not appear to fit well as an emotional response. Excluding lucky from the analyses did not change any of the response patterns in subsequent analyses. Table 5 shows the bivariate correlations

across the remaining 26 responses. We also included ratings of morality in the correlation table, to examine how participants' ratings of the morality of their actions in study correlated with affective responses.

Moral Emotions. To examine whether observation condition and flipping the coin influenced participants' feelings of guilt, pride, shame, and morality regarding their task assignment decision (*prediction 4*), a series of 2 (flip) x 3 (condition) ANOVA's were implemented. Main effects of flip and condition were examined, as well as their interaction. A modified Bonferroni adjustment was utilized to manage family-wise error accumulation, and Type III sums of squares was used in all effect calculations. Of note, ratings of shame, pride, and guilt showed signs of significant positive skew. After normalizing transformations were performed, all analyses were conducted a second time as outlined below. There were no changes in the significance patterns of the results. All reported results are in original (i.e., untransformed) units.

Table 6 shows the descriptive statistics (*M, SD*) for each affective response recorded, by condition and coin flip decision. For feelings of morality, there was a significant effect of flipping the coin, $F(1,155) = 28.19$, $MSE = 5.71$, $p < .001$, $\eta^2_p = 0.15$, with those flipping the coin ($M = 7.05$, $SD = 2.36$) rating themselves as significantly more moral than those who did not ($M = 4.96$, $SD = 2.41$). For feelings of shame, there was a trending effect of flipping the coin, $F(1,155) = 3.71$, $MSE = 1.17$, $p = .056$, $\eta^2_p = .02.$, with those flipping the coin ($M = 1.51$, $SD = 0.87$) rating themselves as less ashamed than those who did not flip the coin ($M = 1.81$, $SD = 1.24$). However, in light of the number of contrasts presented, we emphasize that this result should be interpreted only speculatively.

To examine whether the observation condition and participants' assigning themselves to the positive task influenced their feelings of guilt, pride, shame, and morality regarding their task assignment decision (*prediction 4*), a series of 2 (task assignment) x 3 (condition) ANOVA's were implemented. Main effects of task assignment and condition were examined, as well as their interaction. A modified Bonferroni adjustment was utilized to manage family-wise error accumulation, and Type III sums of squares was used in all effect calculations. Again, normalizing transformations were performed, and all analyses were conducted a second time as outlined below. There were no changes in the significance patterns of the results. All reported results are in original (i.e., untransformed) units.

Table 7 shows the descriptive statistics (*M, SD*) for each affective response recorded, by condition and task assignment decision. Across all emotions, there was a main effect of assigning the other participant to the positive task:

guilt $F(1, 155) = 20.04$, $MSE = 2.52$, $p < .001$, $\eta^2_p = .11$

pride $F(1, 155) = 6.40$, $MSE = 2.95$, $p = .011$, $\eta^2_p = .04$

shame $F(1, 155) = 13.30$, $MSE = 1.10$, $p < .001$, $\eta^2_p = .08$

morality $F(1, 155) = 14.01$, $MSE = 6.22$, $p < .001$, $\eta^2_p = .08$

In other words, across all conditions, participants who assigned the other participant to the positive task felt significantly more moral, more proud, less ashamed, and less guilty, than those who assigned themselves to the positive condition.

We further explored participants' feelings of morality regarding the task assignment by examining the effect of the *outcome* of the coin flip and assignment of the tasks. Coin flip outcome was broken down into three groups: Lost ($M = 6.46$, $SD = 2.63$), Won ($M = 7.55$, $SD = 1.99$), or N/A ($M = 4.96$, $SD = 2.41$), with N/A representing those who did not flip the coin.

Using a 2 (task assignment) x 3 (flip outcome), we found a significant interaction between flip outcome and task assignment, $F(2,155) = 7.74$, $MSE = 4.42$, $p = .001$, $\eta^2_p = .09$. To break down this interaction, we examined the simple main effects of winning at both levels of task assignment, using the global error term for simple main effect calculations. For those who assigned the other person to the positive task, there was no effect of the coin flip outcome, $F(2, 38) = 1.30$, *n.s.* For those who assigned themselves to the positive task, there was an effect of winning, $F(2, 117) = 32.83$, $MSE = 4.42$, $p < .001$. A post-hoc Tukey test revealed that those who won ($M = 7.71$, $SD = 1.75$) rated themselves as significantly more moral than both those who lost ($M = 5.06$, $SD = 2.26$; $p < .001$, $d = 1.31$), and those who did not flip the coin ($M = 4.32$, $SD = 2.20$; $p < .001$, $d = 1.71$). Figure 1 shows the pattern of results for the simple main effects.

Exploratory factor analysis. The bivariate correlations across the emotion responses showed a number of inter-item associations. Subsequently, we entered them into an exploratory factor analysis to examine whether these relationships reflect any underlying factors that may be driving the effect. We used SPSS's principal axis factoring as our extraction method. We first examined the scree plot of factor loadings to help determine an appropriate number of potential underlying factors. We determined that a three factor solution appeared most appropriate, accounting for 54.27 percent of the variance. The Kaiser-Meyer-Olkin measure of sampling adequacy (0.85) and Bartlett's test of sphericity ($\chi^2 = 2384.17$, $p < .001$) indicated our sample size and variable structure was appropriate, given the number of variables entered into the model. Communalities ranged from .253 (happy) to .758 (tender), with the majority falling within the .40 - .60 range.

We used promax oblique rotation with Kaiser normalization for our rotation method, to accommodate potential factor inter-correlations, and because we did not have a strong assumption that the underlying factors would be orthogonal. The rotation converged in 5 iterations. Table 8 shows the full rotated factor matrix, as well as factor inter-correlations. The responses appear to cluster well across the three factors, with only calmness and pride loading onto more than one factor. Based on the loading pattern of the response variables, we interpret these three factors as (1) negative self emotionality, with feelings of unease, unsettled, worried, ashamed, and disturbed loading highly onto this factor; (2) other-centered emotionality, with tender, softhearted, compassionate, warmth, and moved loading highly; and (3) positive self emotionality, with pleased, satisfied, relieved, and happy loading highly.

We calculated mean composite scores for each latent factor, based on the rotated factor loading matrix (see Table 8). We chose to include items with a rotated factor loading above .3, which encompassed all items. For positive emotion we included pleased, satisfied, relieved, and happy; negative emotional included uneasy, unsettled, worried, ashamed, disturbed, guilty, perturbed, afraid, distressed, sad, anxious, irritated, concerned, and calm (reverse coded); other-centered emotion included tender, softhearted, compassionate, warm, moved, sympathetic, and proud. Reliability coefficients showed high internal consistency for all three composite scores (α's = .79, .90, and.88, respectively). We then examined whether these varied across condition, flip decision, and task assignment using a series of two-way ANOVA's, as outlined for the moral emotions analysis.

Table 9 shows the mean level composite scores by condition, coin flip decision, and task assignment decision. Looking at coin flip across conditions, we found a trending effect of coin flip for positive emotion ($F(1,155) = 3.88$, $p = .051$, $\eta^2_p = .02$), such that those who did not

flipped the coin (M = 4.40, SD = 1.13) expressed more positive emotion than those who flipped (M = 4.03, SD = 1.20). We did not find any significant effects for negative emotion or other-centered emotion (negative: $F(1,155) = 0.29$, *n.s.*; other: $F(1,155) = 2.21$, *n.s.*).

Looking at task assignment across conditions, we found a trending effect of task assignment for positive emotion ($F(1,155) = 3.77$, $p = .054$, $\eta^2_p = .02$), such that those who assigned themselves the rewarding task (M = 4.32, SD = 1.10) expressed more positive emotion than those who assigned the other person to the rewarding task (M = 3.91, SD = 1.36). We also found a significant main effect for other-centered emotion ($F(1,155) = 15.38$, $p < .001$, $\eta^2_p = .09$), such that those who assigned themselves the rewarding task (M = 2.90, SD = 1.03) expressed less other-centered emotion than those who assigned the other person to the rewarding task (M = 3,71, SD = 1.32). We did not find any significant effects for negative emotion ($F(1,155) = 1.56$, *n.s.*).

Discussion

We predicted that participants under perceived observation would be more likely to flip the coin (*prediction 1*), and would be less likely to assign themselves to the positive task (*prediction 2*), relative to those who were not. In the present study, we did not find evidence for this effect. The frequency of those who flipped the coin in the low observation condition increased slightly to 65%, and decreased slightly to 36% in the high observation condition (compared to 51% in the no observation condition). However, when these frequencies were directly compared to the no observation condition, participants were not statistically more or less likely to flip the coin. The significant chi-square statistic likely reflected a difference between the low and high conditions, and not a difference between the no observation condition, as predicted by the model. We also predicted that participants who flipped the coin under perceived observation would be more likely use it fairly (*prediction 3*). We did not find this to be the case, and found that task assignment remained surprisingly constant across all conditions, approximating 70% assignment of the self to the positive task.

One explanation for the lack of observer effect is that the mirrors were not sufficient stimuli to tap into the observer tracking mechanisms proposed above. Eye spots have previously been shown to influence behavior in a number of contexts (e.g., Ernest-Jones, Nettle, & Bateson, 2011; Nettle, Nott, & Bateson, 2012; Haley & Fessler, 2005), and while mirrors do produce eye spots (i.e. those of the self), there is little evidence regarding the effectiveness of mirrors as a social observer cue (i.e. strangers' eyes). Thus, as a relatively novel observer stimulus for this effect, it is difficult to tease apart whether mirrors are sufficient to induce an observation effect.

Another possibility is that the consequences for being observed were perceived to be too small to counteract the incentives for assigning the self to the positive task. Consistent with this

explanation, over 70% of participants across all conditions assigned themselves to the positive task. The EU calculations (Fig. 1) take into consideration the costs (*c*) of a transgression when one is observed and punished. Even if social observation cues were present and active, if *c* were very low, the equation would still balance out in favor of taking the rewarding course of action (i.e. assigning the self to the positive task), even when observation likelihood (*p*) is high (such as in the high observation condition). The contrived nature of a laboratory context, the transient one-shot nature of undergraduate research participation, and the institutional trust in researchers at one's university not to induce negative outcomes for research participants, may have aided in the attenuation of *c* as well.

The lack of a mirror effect is in direct contrast to those found by Batson et al. (1999), who provided tentative evidence that a mirror was sufficient to increase moral behavior in this particular dilemma context. Further, this contrasts evidence that mirrors induce a self-awareness effect that can lead to a reduction in antisocial behavior, for example, cheating on a test (Heine et al., 2008; Diener & Wallbom, 1976). However, the relatively small sample size in Batson et al. (1999) raises concerns about the validity of their mirror effect findings. In their study, only 23 participants flipped the coin, and of those only 10 participants flipped the coin under the mirror effect. There also exists limited evidence that the mirror effect does *not* induce self-awareness effects in all situations (e.g., Bögels, Rijsemus & De Jong, 2002). Due to the presence of mixed empirical evidence, more work needs to be done to disentangle the strength and efficacy of mirror-induced awareness effects, as well as exactly what cognitive mechanisms are being activated in these contexts.

We found an effect of gender on task assignment, with females approximately 2.5 times more likely to assign themselves to the positive task than males. As early as 1977, Gilligan

proposed that men and women may have divergent moral foci, with women focusing more on care-based and interpersonal topics and men focusing more on justice-based and rule adherence topics (Gilligan, 1977; more recently, see: Robertson et al., 2007). Since then, research on gender differences in moral reasoning has produced conflicting evidence, with tentative support for Gilligan's initial proposition (for a meta-analytic overview, see: Jaffee & Hyde, 2000). Further, neuroimaging techniques have demonstrated that although men and women display similar evaluative judgments of morally relevant stimuli, they may be engaging different neural circuitry when making these judgments (Harenski, Antonenko, Shane, & Kiehl, 2008). Thus, there may be some theoretical substantiation to this finding in the present study: because men are more sensitive to issues of justice and order, within which fairness could be plausibly nested, men judge the unfair allocation of the tasks to be a stronger moral violation than women. However, we propose this line of reasoning speculatively, and more direct evidence regarding gender differences in moral behavioral settings needs to be collected before a strong conclusion can be made regarding this effect.

We predicted that participants' emotional responses would reflect their behaviors during the task assignment procedure. That is, for the moral emotions we expected those who flipped the coin and assigned the other person the positive task would indicate a higher sense of morality, more pride, less guilt and less shame (*prediction 4*). Relatively consistent with this prediction, when flipping the coin participants felt more moral, but there was no difference in reactions of guilt, pride or shame. When assigning the tasks, participants who assigned the other person to the positive task felt more moral, more proud, less guilty, and less ashamed. Further, we found no significant effect of condition, such that the mirrors appeared to have no effect on the emotional reactions.

Participants' emotional responses appeared to reflect their true behavioral decision: those who assigned the other person to the positive task had nothing to feel guilty or ashamed about, and felt proud of their decision to forego the reward in favor of the other person. On the other hand, those who assigned themselves to the positive task felt guiltier, less proud, and more ashamed of their decision. The simplest explanation is that the emotional responses reflect the proper functioning of the affective system during moral judgment (Haidt, 2003; Tangney et al., 2007). However, this could also be related to the ineffectiveness of the mirrors as a social observer cue, such that the presence of real observers would influence the emotional reactivity of participants. Looking at the trends in affective responding across conditions, there appears to be some consistent patterning. For example, looking at the means in Table 6 we see that guilt and shame are consistently decreasing and pride is consistently increasing, across observation conditions. Thus, perhaps there is an effect but it is too subtle to detect within the large amount of variance present in the self-reported emotions. Further testing would need to be done, ideally with confederate observers and measurements of externalized emotions, to further confirm the effects found in the present study.

Across a total of 23 broader affective reactions, The EFA revealed three distinct underlying factors that we defined as positive self emotion, negative self emotion, and other-centered emotion. Upon further examination of the composite emotion scores across condition, coin flip, and task assignment, we found that positive emotion tended to fluctuate across coin flip and task assignment decision, such that those who did not flip the coin and those who assigned the other person to the positive task expressed more positive self emotion. These appear to be in contrast, as we might have expected those who flipped the coin to feel more positive about their decision. Perhaps this reflects a sense of relief for those who did not flip the coin, or perhaps

those who decided to flip were concerned about the outcome. However, because of the number of comparisons present in the analysis, we emphasize that the effects of positive emotion should be interpreted speculatively.

We also found that other-centered emotion varied across task assignment, such that those who assigned the other person the positive task showed higher scores than those who assigned themselves the positive task. Many of the affective responses measured reflect feelings oriented towards taking an empathic approach to the other person, such as tender, softheared, and compassionate. These appear to reflect concern for the other person, or at the very least, a consideration of the other person during the decision process. Thus, these results appear congruent with what would be expected for individuals who were incorporating thoughts about the other person in the decision process. Although exploratory in nature, these composite scores showed meaningful variation in the current study and may prove a fruitful area for future research.

Individuals showed a salient level of discrepancy between what they considered the most moral course of action to be, and the course of action they actually pursued. This is consistent with previous findings that implemented this particular dilemma (for a concise overview, see Batson et al., 2008). More than 70% of participants claimed that flipping the coin or simply assigning the other person the positive task was the most moral approach to the task assignment procedure. Yet across all conditions more than 70% of participants assigned themselves the positive task, and of those who flipped the coin, 70% of participants assigned themselves to the positive task. This is greater than the 50% we would expect by chance alone, if participants were using the coin flip fairly. Interestingly, of those who flipped and *lost,* and assigned themselves to the positive condition (n = 16), 13 of the 16 (81%) endorsed either flipping the coin or assigning

the other to the positive task as the most moral way to assign the tasks. The other three participants claimed that the decision 'was not morally relevant'. Further, of the 63 participants who decided not to flip the coin and assign themselves the positive task, only 2 people indicated that assigning themselves the positive task was the most moral thing to do.

One possible explanation for these discrepancies might be demand characteristics. Because of the presence of the coin in the study, participants may have felt that we intended for them to use it, leading them to endorse its use as the most moral approach. This could also be perceived as a socially desirable response - flipping the coin, regardless of the outcome, has face validity as a fair and moral approach to the situation. However, there is clear variance in both actions and endorsements, with 50% choosing not to flip the coin, and 44% choosing not to endorse the coin flip as the most moral way to assign the tasks. Thus, although socially desirable responding and demand characteristics may have played a part in some of the participant responding, there appears to be other factors at play.

Alternatively, participants may have been attempting to preserve their reputation (i.e. avoiding costs associated with not acting fairly) while also attaining the rewards associated with the positive task. This is congruent with a moral hypocrisy explanation of moral motivations, as outlined in Batson et al. (2008), where the goal is to appear moral while avoiding the costs of actually following through. Indeed, a self-interested person seeking to maximize their benefits while avoiding costs or missed opportunities would fare best by *endorsing* the socially acceptable or least costly course of action, while *pursuing* that which benefits them the most.

Recent research further supports this explanation. DeScioli et al. (2014) demonstrated that individuals dynamically change their moral judgment of a rule to serve their self-interest. When in a position where an equity-based reward distribution rule (compensation in accordance

with effort expended) was most beneficial for a participant, they were more likely to endorse the equity rule as more fair and moral than the alternative. Conversely, when in a position where an equality-based reward distribution rule (equality of outcome, regardless of effort expended) was most beneficial for the participant, they subsequently endorsed the fairness rule as more fair and moral (DeScioli et al., 2014).

Consistent with this explanation, we found that participants who won the coin flip and assigned themselves to the positive task rated their actions as highly moral. Why would the result of the coin flip influence feelings of morality? Whether the coin flip was won or lost, the act of flipping the coin does not become more or less moral. The outcome of the coin flip influenced how moral participants felt about assigning the tasks, something that would not be expected if participants were simply attempting to pursue the most moral course of action from the outset of the task assignment procedure. Indeed, it appears that those who acted consistently with the coin flip results, win or lose, rated themselves as the most moral, while those who were inconsistent (even those who won and still assigned the other person to the positive task) rated themselves as less moral.

Conclusion

A growing body of evidence is showing that moral judgment is not simply a matter of deontological right and wrong (Broad, 1930; Kant, 1785) but is dependent on the context in which judgment occurs, including but not limited to affective (Valdesolo & DeSteno, 2006), intentional (Greene et al., 2009; Mikhail, 2007), and social (Haley & Fessler, 2005; Rigdon et al., 2009) properties of the situation. Thus, motivation to uphold moral norms *per se* may not be driving decision-making in moral contexts. Socioecological forces, combined with the human desires of the moral agents, may be facilitating this variance across morally relevant situations. As shown in the present study, we found that when posed with a moral dilemma where a person must decide between selfish gains or moral fairness that they will, on average, behave in accordance with a self-interested motivational approach. The discrepancy between what people endorsed as the most moral approach to the situation and how they decided to act provides clear evidence for moral hypocrisy.

In contrast to previous findings, cues of social observation did not appear to influence behavior in a systematic manner, nor did they induce a self-awareness effect. The mirror stimuli used to induce the effect of being observed was an indirect manipulation. This may not have adequately tapped into the proposed observer detection mechanisms as we hoped. As such, we recommend that future investigations of this phenomenon use more direct manipulations of the observer effect, such as images of real people, eyespots, or confederates.

The EU model requires more empirical testing before a firm conclusion can be made about it's viability as a model for moral decision-making. However, EU theory's ability to shape concise, quantifiable, and testable predictions should not be understated. There are a number of empirically testable parameters in the present framework: here we chose to manipulate p, or the

threat of detection and punishment, to see if it would change the likelihood of behavior across groups. Changes in *b* (the rewards or incentives for certain decisions), or changes in *c* (the costs involved for punishment) could also be used to test predictions in moral dilemma situations. As such, we stipulate that more empirical testing should use strong predictive frameworks to quantify and frame decision outcomes in situations with inherent uncertainty.

APPENDIX

Table 1:

Frequency (%) of participant statements of the most moral way to assign the tasks

Statement	Frequency
Flip the coin or another even-handed method	90 (56%)
Assign other to positive task	25 (16%)
Whoever gets the choice can use their discretion	5 (3%)
Involve or communicate with the other person	6 (4%)
Someone else (i.e. the experimenter) should choose	5 (3%)
It depends; whoever needs it most should get it	3 (2%)
The decision is not morally relevant	13 (8%)
Assign self to positive condition	2 (1%)
Other/Unknown/Unspecified	12 (7%)

Table 2:

Number of participants (%) who flipped the coin, and task assignment decision, by condition

	Experimental condition			
Task assignment decision	No mirror	Single mirror	Multiple mirror	Total
Total *N*	55	51	55	161
Coin flip				
Not flip coin	27 (49%)	18 (35%)	35 (64%)	80 (50%)
Flip coin	28 (51%)	33 (65%)	20 (36%)	81 (50%)
Task assignment				
Self to positive task	41 (75%)	36 (71%)	43 (78%)	120 (75%)
Other to positive task	14 (25%)	15 (29%)	12 (22%)	41 (25%)
Task assignment (for only those who flipped the coin)				
Self to positive task	19 (68%)	24 (73%)	14 (70%)	57 (70%)
Other to positive task	9 (32%)	9 (27%)	6 (30%)	24 (30%)

Table 3:

Logistic regression results predicting the likelihood of flipping the coin

Predictor	***df***	***b***	*SE*	***z***	***p***	***OR***	***95% CI***	
(Intercept)		-.066	.37	0.18	.857	0.94	0.45	1.92
Condition	2							
Single mirror		.578	.40	1.45	.148	1.78	0.82	3.95
Multiple mirror		-.592	.39	1.51	.130	0.55	0.25	1.19
Age	1	-.026	.09	0.29	.771	0.97	0.81	1.17
Conservatism	1	-.089	.13	0.69	.491	0.91	0.71	1.18
Gender	1							
Female		.142	.36	0.40	.689	1.15	0.57	2.33

Model statistics

Model χ^2 = 9.42, $p = 0.09$

Pseudo R^2 = 0.08

Table 4:

Logistic regression results predicting the likelihood of assigning self to the positive task

Predictor	***df***	***b***	*SE*	*z*	***p***	***OR***	***95% CI***	
(Intercept)		.749	.44	1.70	.089	2.11	0.91	5.16
Flip		-.459	.39	1.19	.236	0.63	0.29	1.34
Condition	2							
Single mirror		-.146	.45	0.32	.748	0.86	0.35	2.11
Multiple mirror		.075	.47	0.16	.873	1.08	0.43	2.73
Age	1	-.098	.10	1.03	.303	0.91	0.74	1.10
Conservatism	1	.072	.15	0.48	.629	1.08	0.80	1.45
Gender	1							
Female		.940	.39	2.43	.015*	2.56	1.20	5.51

Model statistics

Model χ^2 = 9.18, p = .164

Pseudo R^2 = 0.08

* $p < .05$

Table 5:

Bivariate correlations, affective responses (Pearson's r). Bold coefficients are significant at the .05 level

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.
1. happy	--																									
2. anxious	-.07	--																								
3. sympathetic	-.02	**.32**	--																							
4. concerned	-.13	**.31**	**.28**	--																						
5. softhearted	.04	**.21**	**.59**	**.42**	--																					
6. warm	**.26**	**.19**	**.43**	**.36**	**.67**	--																				
7. distressed	-.07	**.45**	**.28**	**.47**	**.30**	**.30**	--																			
8. compassion	.11	**.28**	**.53**	**.30**	**.67**	**.62**	**.33**	--																		
9. upset	**-.22**	**.39**	**.20**	**.40**	**.23**	**.16**	**.46**	**.25**	--																	
10. tender	.12	**.19**	**.44**	**.42**	**.65**	**.67**	**.35**	**.62**	**.36**	--																
11. moved	.11	**.19**	**.37**	**.32**	**.52**	**.49**	**.33**	**.58**	**.42**	**.77**	--															
12. worried	-.13	**.56**	**.24**	**.44**	**.27**	**.27**	**.48**	**.34**	**.52**	**.34**	**.34**	--														
13. disturbed	-.10	**.28**	**.18**	**.31**	**.19**	**.26**	**.41**	**.22**	**.66**	**.39**	**.48**	**.60**	--													
14. perturbed	-.11	**.30**	**.19**	**.24**	**.18**	**.26**	**.44**	**.23**	**.57**	**.40**	**.47**	**.50**	**.83**	--												
15. uneasy	**-.18**	**.47**	**.27**	**.36**	**.20**	**.19**	**.54**	**.28**	**.64**	**.29**	**.34**	**.66**	**.71**	**.70**	--											
16. relieved	**.29**	**.17**	**.24**	**.17**	**.17**	**.26**	.12	**.19**	.06	**.20**	**.18**	**.28**	.10	.10	**.17**	--										
17. irritated	-.12	**.23**	**.16**	**.18**	.15	**.22**	**.33**	.13	**.43**	**.31**	**.36**	**.23**	**.41**	**.43**	**.45**	-.09	--									
18. sad	-.04	**.33**	.14	**.23**	**.18**	**.20**	**.36**	**.21**	**.62**	**.32**	**.34**	**.37**	**.47**	**.43**	**.42**	.09	**.49**	--								
19. pleased	**.38**	.09	.13	.03	**.18**	**.22**	.12	**.19**	-.08	.11	.08	.15	-.07	-.01	.05	**.59**	-.05	.07	--							
20. guilty	-.07	**.29**	**.28**	**.30**	**.20**	.05	**.36**	.07	**.27**	.06	.00	**.37**	**.27**	**.23**	**.36**	**.20**	**.19**	**.28**	**.25**	--						
21. proud	**.24**	.11	**.22**	.14	**.29**	**.40**	**.18**	**.41**	**.21**	**.42**	**.43**	.13	.11	.15	.14	**.37**	**.22**	**.19**	**.42**	-.06	--					
22. afraid	.03	**.30**	**.16**	**.23**	.13	**.16**	**.27**	**.25**	**.41**	**.28**	**.36**	**.45**	**.44**	**.43**	**.44**	.13	**.39**	**.43**	.04	**.29**	.13	--				
23. ashamed	.02	**.25**	**.16**	**.28**	.09	**.16**	**.35**	.07	**.37**	**.19**	.15	**.45**	**.39**	**.36**	**.42**	.13	**.41**	**.39**	.10	**.62**	.02	**.48**	--			
24. satisfied	**.40**	.02	.10	.04	**.16**	**.17**	-.02	**.18**	-.07	**.17**	.15	.10	.01	-.01	-.01	**.60**	**-.19**	.01	**.62**	.05	**.54**	.01	-.05	--		
25. unsettled	-.15	**.37**	.09	**.39**	.10	.15	**.44**	.06	**.48**	**.21**	**.16**	**.49**	**.52**	**.50**	**.64**	.11	**.39**	**.37**	.01	**.45**	-.04	**.42**	**.54**	-.09	--	
26. calm	**.27**	**-.35**	-.06	-.14	.04	.10	**-.27**	-.07	**-.30**	.01	-.03	**-.29**	**-.27**	**-.19**	**-.31**	**.19**	-.15	**-.24**	**.27**	**-.20**	**.25**	**-.30**	**-.20**	**.38**	**-.24**	--
27. moral	**-.17**	.02	**.26**	.05	**.25**	**.26**	.06	**.31**	.03	**.17**	.15	-.02	.05	.04	-.02	-.03	.03	-.03	-.07	**-.19**	.11	-.02	**-.26**	.01	**-.24**	-.03

Table 6:

Mean (SD) affective ratings after task assignment decision, by condition and flip

	Experimental condition			
	No mirror	Single mirror	Multiple mirror	Total
Flip				
Moral	6.57 (2.54)	7.15 (2.29)	7.55 (2.16)	7.05 (2.36)
Guilt	2.54 (1.90)	2.52 (1.46)	2.15 (1.50)	2.43 (1.62)
Pride	3.00 (1.68)	2.91 (1.76)	2.70 (1.81)	2.89 (1.72)
Shame	1.57 (0.84)	1.45 (0.75)	1.50 (1.10)	1.51 (0.87)
No flip				
Moral	5.07 (2.53)	5.33 (2.81)	4.69 (2.13)	4.96 (2.41)
Guilt	2.48 (1.76)	2.44 (1.85)	2.37 (1.66)	2.43 (1.72)
Pride	3.52 (1.91)	3.33 (1.94)	2.94 (1.49)	3.23 (1.74)
Shame	1.74 (1.40)	2.06 (1.30)	1.74 (1.09)	1.81 (1.24)

Table 7:

Mean (SD) affective ratings after task assignment decision, by task assignment and flip

	Experimental condition			
	No mirror	Single mirror	Multiple mirror	Total
Self to Positive				
Moral	5.32 (2.47)	6.25 (2.66)	5.26 (2.56)	5.58 (2.58)
Guilt	2.80 (1.93)	2.83 (1.66)	2.63 (1.65)	2.75 (1.74)
Pride	3.00 (1.75)	2.81 (1.60)	2.74 (1.59)	2.85 (1.64)
Shame	1.78 (1.27)	1.89 (1.12)	1.83 (1.17)	1.83 (1.18)
Other to Positive				
Moral	7.36 (2.53)	7.13 (2.47)	7.42 (1.62)	7.29 (2.23)
Guilt	1.64 (1.08)	1.67 (1.05)	1.08 (0.29)	1.49 (0.93)
Pride	4.00 (1.80)	3.67 (2.19)	3.25 (1.66)	3.66 (1.89)
Shame	1.29 (0.47)	1.13 (0.35)	1.00 (0.00)	1.15 (0.36)

Table 8:

R*otated factor loadings and factor inter-correlation matrix. Only factor loadings >.3 displayed*

	Factor 1	Factor 2	Factor 3
uneasy	.832		
unsettled	.814		
worried	.723		
ashamed	.710		
disturbed	.701		
upset	.675		
guilty	.648		
perturbed	.645		
afraid	.573		
distressed	.562		
sad	.558		
anxious	.537		
irritated	.440		
calm	-.420		.363
concerned	.361		
tender		.866	
softhearted		.796	
compassionate		.795	
warm		.755	
moved		.754	
sympathetic		.508	
proud		.441	.389
pleased			.821
satisfied			.802
relieved			.726
happy			.437
	Factor 1	Factor 2	Factor 3
Factor 1	--		
Factor 2	.42	--	
Factor 3	-.07	.211	--

Table 9:

Mean (SD) composite scores across conditions, by coin flip and task assignment decisions

	Experimental condition			
	No mirror	Single mirror	Multiple mirror	Total
Other to positive				
positive	4.49 (1.25)	4.17 (0.91)	4.28 (1.09)	4.32 (1.10)
negative	2.19 (0.93)	2.30 (0.82)	2.33 (0.98)	2.27 (0.87)
other	2.74 (0.92)	3.15 (1.09)	2.83 (1.06)	2.90 (1.03)
Self to positive				
positive	4.34 (1.40)	3.82 (1.41)	3.54 (1.23)	3.91 (1.36)
negative	2.05 (0.72)	2.18 (0.74)	1.99 (0.66)	2.08 (0.70)
other	4.04 (1.20)	3.15 (1.09)	3.32 (1.06)	3.71 (1.32)
No flip				
positive	4.34 (1.40)	3.82 (1.41)	3.54 (1.23)	3.91 (1.36)
negative	2.05 (0.72)	2.18 (0.74)	1.99 (0.66)	2.08 (0.70)
other	4.04 (1.20)	3.71 (1.59)	3.32 (1.06)	3.71 (1.32)
Flip				
positive	4.48 (1.25)	4.17 (0.91)	4.28 (1.09)	4.32 (1.10)
negative	2.19 (0.93)	2.30 (0.82)	2.33 (0.98)	2.27 (0.92)
other	2.74 (0.92)	3.15 (1.09)	2.83 (1.06)	2.90 (1.03)

Note: positive = positive emotion; negative = negative emotion; other = other-centered emotion

Figure 1:

Expected utility model parameters of moral behavior

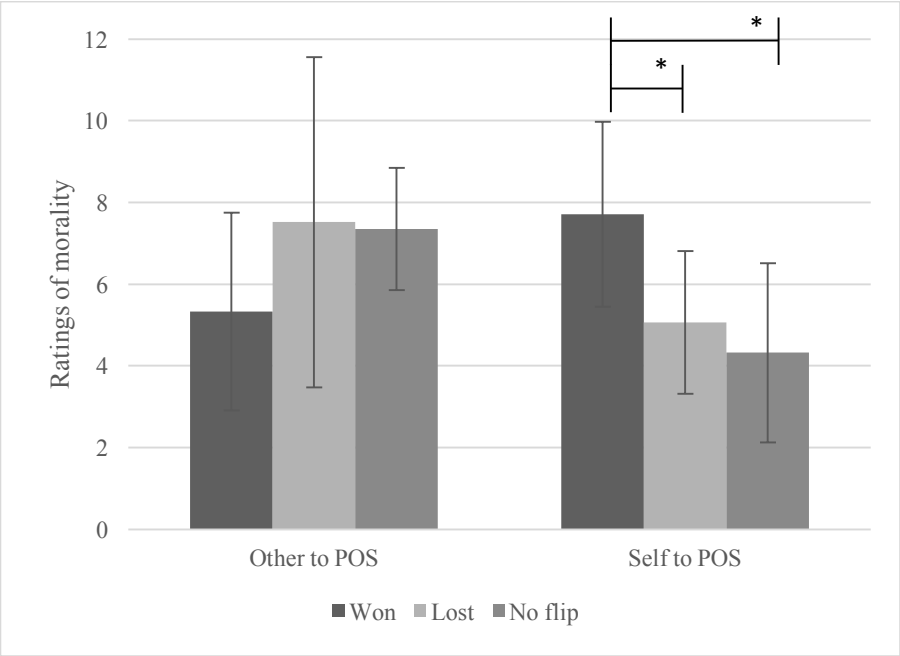
	Caught (p)	Not caught (1-p)
Action	c	b
No action	0	0

EU(Action) = $c(p) + b(1-p)$

EU(No action) = 0

Figure 2:

Ratings of morality by coin flip outcome (loss, win, or no flip) by task assignment (self to positive, other to positive)



Note: *$p < .05$; error bars denote standard deviation.

REFERENCES

REFERENCES

Batson, C. (2008). Moral masquerades: Expermental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences, 7*(1), 51-66. doi: 10.1007/s11097-007-9058-y

Batson, C., Kobrynowicz, D., Dinnerstein, J., Kampf, H., & Wilson, A. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348. doi: 10.1037/0022-3514.72.6.1335

Batson, C., Thompson, E., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology, 83*(2), 330-339. doi: 10.1037//0022-3514.83.2.330

Batson, C., Thompson, E., Seuferling, G., Whitney, H., & Strongman, J. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525-537. doi: 10.1037/0022-3514.77.3.525

Bloom, P. (2013). *Just Babies: The Origins of Good and Evil* (1st ed.). New York, NY: Crown Publishing Group.

Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York, NY: Basic Books.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. (2003). The evolution of altruistic punishment. *PNAS, 100*(6), 3531-3535. doi:10.1073/pnas.0630443100

Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*(3), 171-195. doi: 10.1016/0162-3095(92)90032-Y

Broad, C. (1930). *Five types of ethical theory*. New York: Harcourt, Brace and Co.

Burnham, T., & Hare, B. (2007). Engineering human cooperation: does involuntary neural activation increase public goods contributions? *Human Nature, 18*(2), 88-108. doi: 10.1007/s12110-007-9012-2

Cosmides, L., & Tooby, J. (2005). *Neurocognitive adaptations designed for social exchange*. In D. Buss (Ed.), The Handbook of Evolutionary Psychology (pp. 584-627). Hoboken, NJ: Wiley.

Cosmides, L., Tooby, J., Fiddick, L., & Bryant, G. (2005). Detecting cheaters. *Trends in Cognitive Sciences, 9*(11), 505-506. doi: 10.1016/j.tics.2005.09.005

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition, 127*(1), 6-21. doi: 10.1016/j.cognition.2012.11.008

DeScioli, P., Massenkoff, M., Shaw, A., Petersen, M., & Kurzban, R. (2014). Equity or equality? Moral judgments follow the money. *Proceedings of the Royal Society B, 281*(1797), 20142112. doi:10.1098/rspb.2014.2112

DeScioli, P., Asao, K., & Kurzban, R. (2012). Omissions and byproducts across moral domains. *PLoS ONE, 7(*10), e46963. doi: 10.1371/journal.pone.0046963

DeScioli, P., Christner, J., & Kurzban, R. (2011). The omission strategy. *Psychological Science, 22*(4), 442-446. doi: 10.1177/0956797611400616

Diener, E., & Wallbom, M. (1976). Effects of self-awareness on antinormative behavior. *Journal of Research in Personality, 10*(1), 107-111. doi:10.1016/0092-6566(76)90088-X

Ekman, P. (2007). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life* (2nd ed.). New York, NY: Owl Books.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 376-379. doi:10.1037/0003-066X.48.4.384

Enoch, J. (2006). History of mirrors dating back 8000 years. *Optometry and Vision Science, 83*(10), 775-781. doi:1040-5488/06/8310-0775/0

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior, 32*(3), 172-178. doi: 10.1016/j.evolhumbehav.2010.10.006

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87. doi: 10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137-140. doi: 10.1038/415137a

Fishburn, P. (1982). The foundations of expected utility. Boston, Dordrecht and London: Reidel.

Friedman, H., & Miller-Herringer, T. (1991). Nonverbal display of emotion in public and in private: Self-monitoring, personality, and expressive cues. *Journal of Personality and Social Psychology, 61*(5), 766-775. doi:10.1037/0022-3514.61.5.766

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard Educational Review, 47*(4).

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., & Cohen, J. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364-371. doi: 10.1016/j.cognition.2009.02.001

Haidt, J. (2003). *The moral emotions*. In R. Davidson, K. Scherer & H. Goldsmith (Eds.), Handbook of affective sciences (pp. 852-870). Oxford: Oxford University Press.

Haley, K., & Fessler, D. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior, 26*(3), 245-256. doi: 10.1016/j.evolhumbehav.2005.01.002

Harenski, C., Antonenko, O., Shane, M., & Kiehl, K. (2008). Gender differences in neural mechanisms underlying moral sensitivity. *Social Cognitive and Affective Neuroscience, 3*(4), 313-321. doi:10.1093/scan/nsn026

Haselton, M., & Buss, D. (2000). Error Management Theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*(1), 81-91. doi: 10,1037110022-3514.78.1.81

Heine, S., Takemoto, T., Moskalenko, S., Lasaleta, J., & Henrich, J. (2008). Mirrors in the head: Cultural variation in objective self-awareness. *Personality and Social Psychology Bulletin, 34*(7), 879-887. doi:10.1177/0146167208316921

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770. doi:10.1126/science.1127333

Jaffee, S., & Hyde, J. (2000). Gender differences in moral orientation: A meta-analysis. *Psychological Bulletin, 126*(5), 703-726. doi:10.1037/TO33-2909.126.5.703

Johnson, D., Blumstein, D., Fowler, J., & Haselton, M. (2013). The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28*(8), 474-481. doi: 10.1016/j.tree.2013.05.014

Kant, I. (1785). *First section: Transition from the common rational knowledge of morals to the philosophical Groundwork of the Metaphysics of Morals.*

Mascolo, M., & Fischer, K. (1995). Developmental transformations in appraisals for pride, shame, and guilt. In J. Tangney & K. Fischer (Eds.), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride.* (pp. 64-113). New York, NY: Guilford Press.

Mikhail, J. (2007). Universal moral grammar: theory, evidence, and the future. *Trends in Cognitive Sciences, 11*(4), 143-152. doi: 10.1016/j.tics.2006.12.007

Monin, B., & Merritt, A. (2011). *Moral hypocrisy, moral inconsistency, and the struggle for moral integrity*. In M. Mikulincer & P. R. Shaver (Eds.), The social psychology of morality: Exploring the causes of good and evil, Herzliya Series on Personality and Social Psychology (Vol. 3). Washington, DC: American Psychological Association.

Nettle, D., Nott, K., & Bateson, M. (2012). 'Cycle thieves, we are watching you': Impact of a simple signage intervention against bicycle theft. *PLoS ONE, 7*(12), e51738. doi:10.1371/journal.pone.0051738

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology, 30*(3), 358-367. doi: 10.1016/j.joep.2009.02.002

Robertson, D., Snarey, J., Ousley, O., Harenski, K., Bowman, F., Gilkey, R., & Kilts, C. (2007). The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia, 45*(4), 755-766. doi:10.1016/j.neuropsychologia.2006.08.014

Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science, 17*(6), 476-477. doi: 0.1111/j.1467-9280.2006.01731.x

Wicklund, R. (1975). Objective self-awareness. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8, pp. 233-275). New York, NY: Academic Press.

Yu, J., Tseng, P., Muggleton, N., & Juan, C. (2015). Being watched by others eliminates the effect of emotional arousal on inhibitory control. *Frontiers in Psychology, 6*(4), 1-5. doi:10.3389/fpsyg.2015.00004