NOW IN PRESS AT CONSCIOUSNESS AND COGNITION (AS OF MAY 15, 2018)

Are morally good actions ever free?

Cory J. Clark

Florida State University

Adam Shniderman

University of Michigan

Jamie B. Luguri

University of Chicago

Roy F. Baumeister

Florida State University and University of Queensland

Peter H. Ditto

University of California, Irvine

Please contact Cory J. Clark with any questions concerning this article at: Department of Psychology, Florida State University. 1107 W. Call Street, Tallahassee, FL. 32306-4301. Tel: (330) 714-9094, Email: callcoryclark@gmail.com

The work of Cory J. Clark was supported in part by a grant from The Charles Koch Foundation.

Abstract

Research has shown that people ascribe more responsibility to morally bad actions than both morally good and neutral ones, suggesting that people do not attribute responsibility to morally good actions. The present work demonstrates that this is not so: People ascribe more free will to morally good than neutral actions (Studies 1a-1b, Mini Meta). Studies 2a-2b distinguished the underlying motives for ascribing freedom to morally good and bad actions. Free will ascriptions for immoral actions were driven predominantly by affective responses (i.e., punitive desires, moral outrage, and perceived severity of the crime). Free will judgments for morally good actions were similarly driven by affective responses (i.e., reward desires, moral uplift, and perceived generosity), but also more pragmatic considerations (perceived utility of reward, counternormativity of the action, and required willpower). Morally good actions may be more carefully considered, leading to generally weaker, but more contextually sensitive free will judgments.

Keywords: free will, morality, praise, blame, motivated cognition, affect, punishment, reward, responsibility

Are Morally Good Actions Ever Free?

1. Introduction

In October 2015, the Concord, Massachusetts police department began issuing citations to their residents for behaviors such as crossing the street at marked crosswalks, wearing seatbelts, and yielding to pedestrians. These citations did not bring fines or jail time but rather could be redeemed for ice cream at a local café. The peculiarity of these good-behavior citations gained the community media attention. However, the program was not adopted as a permanent policy for incentivizing law-abiding behavior within the community. Whereas most governments have penal systems for punishing rule breakers, very few have systems for rewarding rule followers or philanthropists.

On a societal level, establishing accountability for harmful actions and outcomes is highly prioritized above establishing accountability for helpful actions and outcomes. This same moral valence asymmetry has been found repeatedly in individual judgments of accountability. People ascribe more responsibility to morally bad actions and outcomes than morally good ones (e.g., Alicke, 1992; Knobe, 2003; Reeder & Spores, 1983). Expanding on this work, recent research has demonstrated that "bad is freer than good" (p.26): People attribute more free will to bad actions and actions with bad outcomes than good actions and actions with good outcomes (Feldman, Wong, & Baumeister, 2016). These and many similar results seem to suggest that people are disinclined to assign responsibility for good actions. To our knowledge, however, no work has compared responsibility judgments for morally good actions to morally neutral ones. Because rewarding prosocial behavior, like punishing antisocial behavior, is beneficial for group functioning, we hypothesized that reward and praise motives might also increase ascriptions of free will. Therefore, we expected that people would ascribe more free will to morally good

actions than morally neutral ones, but, replicating past work, free will judgments would be higher for morally bad actions than both morally good and neutral actions.

The present research also sought to test the hypothesis that responses to morally good actions would be more context-sensitive than responses to bad actions. Outrage, indignation, and other emotional responses to bad actions produce strong impulses to punish, regardless of whether the punishment will have any deterrent effect (e.g., Carlsmith, Darley & Robinson, 2002; Crockett, Özdemir, & Fehr, 2014). These impulses motivate individuals to perceive harmful actions as freely performed, which helps to justify punishing the wrongdoers (e.g., Clark et al., 2014). In contrast, positive emotional responses to morally positive actions may increase sensitivity to contextual features of the act, consistent with the generally broadening effect of positive emotion (e.g., Frederickson, 2001, 2013; Frederickson & Branigan, 2005). When people witness morally good actions, they may more carefully evaluate whether the behavior is deserving of praise and whether rewarding the behavior would encourage future good behavior. We hypothesized that free will judgments for morally bad actions would be driven predominantly by affective punitive motives, whereas people might more carefully consider a variety of relevant features when determining whether morally good actions were freely performed.

1.1 The Good, The Bad, and The Neutral

Psychology has consistently shown that negativity plays a greater role in peoples' lives than positivity (Kanouse & Hanson, 1971; Lewicka, Czapinski, & Peeters, 1992; Skowronski & Carlston, 1989). A sweeping literature review concluded that "bad is stronger than good" (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001): people give greater weight to negative events, objects, and personality traits than good ones (Rozin & Royzman, 2001). Regarding

responsibility-related judgments for morally relevant behavior, myriad studies have demonstrated that people attribute more responsibility to bad actions and actions with bad outcomes than to closely matched 1) good actions and actions with good outcomes (e.g., Alicke, 1992; Knobe, 2003; Reeder & Spores, 1983), 2) neutral actions and actions with neutral outcomes (e.g., Cushman, Knobe, & Sinnott-Armstrong, 2008; Knobe & Fraser, 2008), and 3) less bad actions and actions with less bad outcomes (e.g., Walster, 1966). Most relevant to the present work, people attribute more free will to morally bad actions and outcomes than to morally good (e.g., Everett et al., 2017) and morally neutral ones (Clark et al., 2014).

This body of work seems to suggest that people do not attribute responsibility to morally good actions, but very little work has compared responsibility judgments for morally positive actions to morally neutral ones. In one cross-cultural analysis, participants from the U.S. and India attributed less intentionality to actions with helpful side-effects than neutral side-effects, seeming to suggest that people do not attribute responsibility to helpful actions (Clark, Bauman, Kamble, & Knowles, 2016). However, in these same studies, praise motives were strong predictors of intentionality judgments for actions with helpful side-effects, seeming to suggest that praise motives do increase attributions of responsibility. To our knowledge, no similar work has compared responsibility judgments between morally positive and neutral actions where the moral valence of the action was not a mere side-effect. Thus, it remains unknown whether praise motives elicit heightened responsibility judgments for morally positive actions relative to morally neutral ones. The first goal of the present research was to investigate these differences in the context of free will judgments. We sought to examine whether exposure to morally positive actions, and subsequent motives to praise, would increase free will judgments. We readily concede that bad actions and subsequent punitive motives elicit the strongest free will judgments

— but good actions and praise motives may elicit stronger free will judgments than morally neutral actions.

1.2 Free Will Judgments

At least since the days of Aristotle, philosophers and other intellectuals have debated the existence of human free will, a debate revived in recent years by neuroscientists and psychologists (e.g., Bargh, 2008; Baumeister, Clark, & Luguri, 2015; Libet, 1985; Soon, Brass, Heinze, & Haynes, 2008; Wegner & Wheatley, 1999). Despite a persistent lack of consensus among the philosophical and scientific communities on the issue, members of the general public believe strongly in free will, across ages and cultures (Nahmias, Morris, Nadelhoffer, & Turner, 2005; Nichols, 2004; Sarkissian et al., 2010). Furthermore, folk beliefs in this concept are relevant to a variety of real-world moral consequences. For example, an experimental manipulation that reduces free will beliefs causes an increase in antisocial behavior (Baumeister, Masicampo & DeWall, 2009; Protzko, Ouimette, & Schooler, 2016; Vohs & Schooler, 2008) and a decrease in punitiveness (Shariff et al., 2014). Performing antisocial acts and declining to punish others who misbehave both can weaken the moral consensus that helps society benefit its members.

1.2.1 Motivated free will judgments. Many factors contribute to free will beliefs, including the powerful subjective experience of conscious will (Wegner, 2002, 2003) and observations about causality (Nichols, 2004). However, one driving factor is a strong motive to justify punishing immoral behaviors (Clark, Baumeister, & Ditto, 2017). In five studies, Clark and colleagues (2014) demonstrated that exposure to the morally bad actions of others, and subsequent desires to punish immoral actions lead people to attribute more free will to such actions and to believe more in the general human capacity for free action. These findings

provided empirical support for Nietzsche's (1889, 1954) contention that the concept of free will was invented to satisfy instincts to judge and punish others.

When it comes to moral responsibility judgments, the general consensus is that people ought to be punished only for actions that they freely chose (Nichols & Knobe, 2007). Accordingly, when people desire to punish others for harmful actions, they increase their perceptions that the misdeeds were performed freely. In other words, desires to hold others morally responsible (particularly for immoral actions) influence judgments about the very factor necessary to warrant moral responsibility: whether the person performed the action of their own free will. Although people generally regard their judgments as based on logical reasoning, preferences for certain conclusions often influence the judgments people make and the beliefs they hold (e.g., Baumeister & Newman, 1994; Ditto & Lopez, 1992; Kunda, 1990). Moral judgments in particular are less often characterized by rigorous contemplation of moral principles than by affective and intuitive reactions to potentially morally relevant stimuli (Clark, Chen, & Ditto, 2015; Ditto, Pizarro, & Tannenbaum, 2009). Instead of preceding moral judgment, moral reasoning often follows as a post-hoc justification for one's desired moral conclusions (Haidt, 2001; Liu & Ditto, 2013). In the case of free will judgments, desires to blame and punish immoral actions lead people to perceive such actions as freely performed, thus providing a crucial justification for punishing such actions (e.g., Alicke, 2000; Clark et al., 2017).

One analysis of these motivated free will judgments suggests that attributing free will to actors signifies that the actor is accountable for the outcome and capable of change or learning to produce better outcomes in the future (Feldman et al., 2016). Consistent with that view, people perceive others as incapable of learning in a universe without free will (Feldman &

Chandrashekar, 2017). Much like punishment is an effective tool for teaching others to cooperate (Cushman, 2013), ascribing free will to an individual may lead to feelings of personal accountability, which could encourage individuals to learn from past mistakes (Alquist, Ainsworth, Baumeister, Daly, & Stillman, 2015). Thus in addition to justifying punitive responses, ascribing free will to immoral actors may effectively promote more desirable behavior in the future.

1.3 Retributive Punishment

In principle, punishment could be a purely dispassionate, pragmatic response, engaged in only for the sake of deterring bad behavior—or it could be affectively motivated. Although punishment does effectively deter antisocial behavior, and people express support for deterrent punitiveness (e.g., Crockett et al., 2014), the actual motivations underlying punishment appear predominantly retributive (e.g., Carlsmith, 2008; Carlsmith et al., 2002). For example, Carlsmith (2006) found that when determining how much to punish someone, people prefer information related to retribution (e.g., the severity of the harm) over information related to the likely utility of the punishment (e.g., whether the punishment will become public knowledge). Other work has demonstrated that people will punish unfair others even when there will be no deterrent benefits (Crockett et al., 2014). Punishing any particular instance of harmful behavior may effectively deter that person from reoffending and deter others from engaging in similar harmful behaviors, however, this does not appear to be the proximate reason humans punish one another. Rather, humans *want* to punish harmdoers, even when the harm is directed toward unknown others, and even when punishing comes at a personal cost (Fehr & Gächter, 2002; Henrich et al., 2005; Henrich et al., 2006). Exposure to another's harmful action elicits a strong affective response, to wit, anger, which leads people to punish (e.g., Fehr & Fischbacher, 2004; Nelissen &

Zeelenberg, 2009). These strong affective responses lead people to punish regardless of potentially relevant characteristics of the harmful act and the punishment, such as whether punishing the harmful action will have any utility.

Consider the concepts "blinded by rage" and "crime of passion." The implications here are that anger and desires for revenge can crowd out rational considerations and limit attention to other important features of a situation. These strong punitive tendencies play an integral role in society (e.g., Cushman, 2013; Fehr & Gächter, 2005). Unchecked human behavior is often selfish (e.g., Karau & Williams, 1993; Kerr & Bruun, 1983; Latané, Williams, & Harkins, 1979), and instances of selfish behavior can lead others to defect to the detriment of entire social groups (Kerr, 1983; Orbell & Dawes, 1981). Punishment serves the crucial social function of deterring selfish behavior (e.g., Fehr, Gächter, Kirchsteiger, 1997), and thus punishing rule breakers is likely an indispensable feature of human cultures (e.g., Clutton-Brock & Parker, 1995; Henrich et al., 2010).

1.4 Punishment is Stronger than Praise

With respect to influencing future behavior, blame and punishment appear to be more impactful than praise and reward. For example, Rasmussen and Newland (2008) found that punishment had a greater impact on an individual's future behavior than reinforcement of equivalent magnitude. This asymmetry has been found in a number of dimensions related to moral judgments, including evaluations of character, control, and intentionality. For example, people are hypersensitive to moral violations, even at the expense of failing to recognize cases of virtuous action (Monroe, Ainsworth, Vohs, & Baumeister, 2017). In fact, some analyses suggest that rewarding prosocial behavior will sometimes undermine intrinsic motivation to help others (Deci, Koestner, & Ryan, 1999; Fabes, Fultz, Eisenberg, May-Plumlee, & Christopher, 1989;

Warneken & Tomasello, 2008, though also see Cameron, Banko & Pierce, 2001). Hence it seems far more useful to punish bad behavior than to reward good behavior.

Nonetheless, both punishment and reward do seem to effectively mold future behavior (e.g., Kubanek, Snyder, & Abrams 2015). Just as crime and the resulting punishment are a normal part of society and play a significant role in defining socially acceptable behavior, prosocial behavior, ranging from basic cooperation to altruism, and the recognition and rewarding of that behavior contribute to successful group functioning (Irwin, 2009). Much like punishment, reward can promote prosocial behavior—at least in some circumstances (Angrist & Lavy, 2009; Ariely, Bracha, & Meier, 2007; Croson & Gneezy, 2009; Gneezy, Meier, & Rey-Biel, 2011; Lacetera & Macis, 2010, Szolnoki & Perc, 2010). One study found that people reward positive behavior approximately as often as they punish negative behavior, and furthermore, that reward and punishment both increase net payoffs, suggesting that both punishment and reward are important for optimal group functioning (Fehr et al., 1997). Due to the value of rewarding morally good behavior, people may be motivated to see good actions as freely performed and good actors as morally responsible.

1.5 Affective Responses to Morally Positive Behavior

As with punishment, reward could be motivated by either affectively charged feelings or by dispassionate calculation of what will benefit the self and society. That is, people may reward good behavior because they are compelled by positive emotional responses (like anger responses lead to retributive punitive desires) — or they might confer rewards based on the notion that rewarding virtuous behavior increases such behavior, thereby making society better for its members (i.e., utilitarian concerns). Only recently have researchers begun investigating affective responses to morally virtuous acts. The positive emotional response triggered by exposure to

morally good acts has been termed *elevation*, which is associated with feelings of being uplifted and increased desires to behave prosocially toward others (Haidt, 2003).

To our knowledge, no work has attempted to tease apart affective vs. more utilitarian motives to praise or reward morally good behavior (as has been done with retributive vs. deterrent punitiveness). There is, however, some reason to expect that praise and reward responses would be more sensitive to utilitarian concerns than punitive motives. The Broaden-and-Build Theory of positive emotions argues that positive emotions broaden the scope of attention and expand awareness, thereby allowing people to consider more contextual information (e.g., Frederickson, 2001, 2013; Frederickson & Branigan, 2005) and to incorporate more relevant information into their decision-making, as compared to negative emotional states (Isen, Rosenzweig, & Young, 1991). To our knowledge, there are no analogous concepts to "blind rage" to describe relatively thoughtless and insensitive desires to praise or reward. A recent literature review concluded that elevation actually expands awareness and flexible thinking (Pohling & Diessner, 2016). Therefore, when evaluating moral responsibility for morally positive actions, people may consider more contextual information, such as whether rewarding positive behavior will encourage future good behavior.

1.6 The Present Research

When people are exposed to morally bad actions, they may be overwhelmed by strong affective desires to punish, which renders their judgments insensitive to various features of the act itself as well as features of the punishment (such as whether there will be any deterrent benefits; e.g,. Crockett et al., 2014). When people are exposed to morally good actions, positive feelings may broaden attention to contextual features of the specific act, allowing for more critical evaluation and consideration of the likely utility of reward. If so, people might ascribe

free will to morally bad acts regardless of whether punishment would have deterrent effects, due to an overwhelming urge to hold wrongdoers morally responsible. But when people evaluate morally good acts, their free will ascriptions might be sensitive to a variety of relevant features, including whether rewards would encourage morally good behavior in the future.

The present work tested the hypothesis that people perceive morally positive actions as more freely performed than morally neutral actions. We acknowledge that there are stronger motives to interpret bad actions as freely performed than there are for morally good actions—but exposure to good actions could also produce a non-negligible increment in perceived freedom. Hence morally positive actions (relative to more neutral actions) might also inspire motives to perceive moral actors, actions, and mankind, as free. This was tested in Studies 1a, 1b, and a mini-meta-analysis of the two studies. Studies 2a and 2b moved on to explore potential reasons for this asymmetry by testing the hypothesis that free will judgments for immoral actions are driven predominantly by affective punitive motives, whereas free will judgments for morally positive actions are more sensitive to a variety of contextual features of morally good acts, in particular, whether rewards will have utility.

2. Study 1a

Study 1a provided a first test of the hypothesis that people would believe more in free will after considering morally good actions than morally neutral ones. Participants read and evaluated an anti-free will research passage. Embedded in that passage were remarks from a fictional academic about the implications of the research, which indicated that people are not responsible for their everyday, mundane behavior (neutral condition), should not be praised for their good behavior, or should not be punished for their bad behavior. We predicted that people would most critically evaluate the anti-free will research in the morally bad condition and least

critically in the neutral condition, with the morally good condition falling between these two. This would indicate that people desire to uphold the idea of free will the most when they wish to punish bad behavior, but also to a (lesser) degree when they wish to reward morally virtuous behavior.

2.1 Method

Participants (n = 234) were recruited through Mechanical Turk (MTurk). Procedures were adapted from Scurich and Shniderman (2014). The target sample size was 240 (~80/condition, based on the experimenter's experience conducting similar research);[1] however research funds ran out after 234 participants. After reporting demographics, participants completed a standard measure of free will belief, the free will subscale of the Free Will and Determinism scale (FAD-Plus; Paulhus & Carey, 2011), on which people report their agreement with a variety of statements such as "People have complete control over the decisions they make" on a 5-point scale from *Strongly disagree* to *Strongly agree*, α = .88. Participants were then randomly assigned to read and evaluate one of three newspaper articles, which described recent scientific research opposing the existence of free will. In the morally bad condition, the lead researcher commented that these results indicate that people should not be held responsible and punished for their bad behavior. In the morally good condition, the researcher stated that people should not be praised for their altruistic or morally good behavior. In the neutral condition, the researcher stated that individuals are not responsible for everyday choices like what clothes they wear or what route they take to work. Exact phrasing of the materials for each study is available in the supplementary materials.

[1] Adam Shniderman was the experimenter for this study. Scurich and Shniderman (2014) recruited approximately 85 participants per condition and ended up with approximately 75 participants per condition in each of their two studies.

Following the newspaper article, participants evaluated the scientific quality of the research on 11 questions, which were presented in randomized order (e.g., "Do you believe the results of the study are likely to be replicable in the future?" "How likely is it that most scientists would agree with the researchers' conclusions?"), on relevant 7-point scales, $\alpha = .90$. Because people more critically evaluate information that opposes their beliefs than information that supports their beliefs (e.g, Munro & Ditto, 1997), a person's beliefs can be inferred from his or her evaluation of information. More critical evaluations of the anti-free will argument would indicate stronger belief in free will. This same method for assessing free will beliefs was also used in Clark et al. (2014), as this more indirect measurement approach reduces the likelihood of demand characteristics.

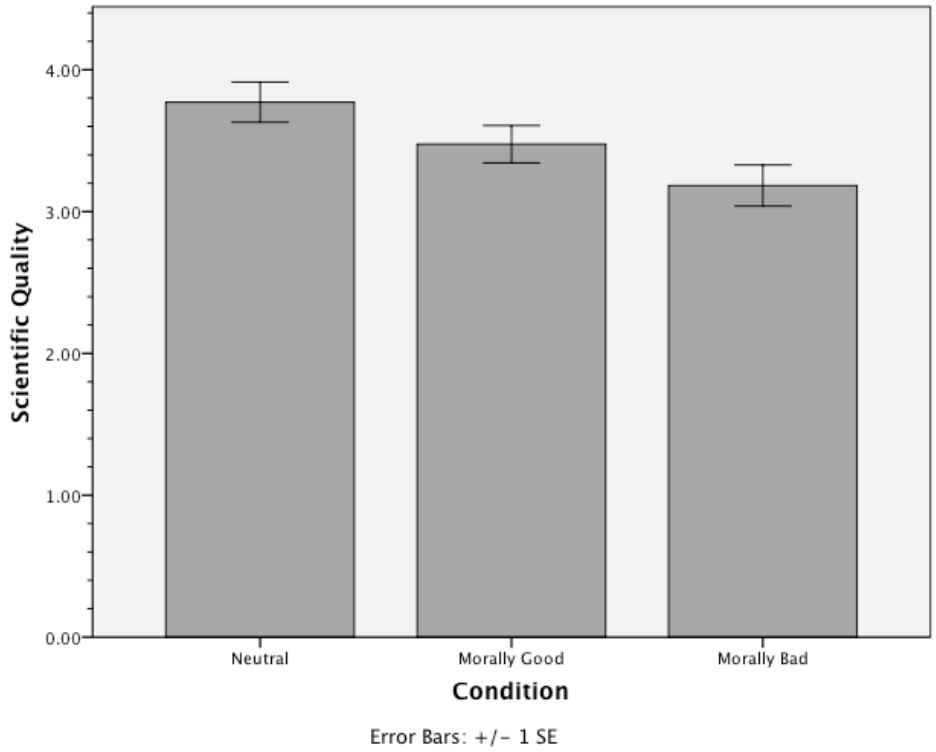
As comprehension checks, participants were asked to identify the technique used in the fictional study (i.e., neuroscience) and to identify the topic of the news article (i.e., free will). Participants also completed a simple attention check that stated, "To ensure this survey is working properly, please click *Strongly Agree*." Participants who failed the comprehension checks ($n = 20$) or attention check ($n = 7$) were removed prior to analyses, resulting in a final sample of 207 participants (98 female; $M_{age} = 35.30$). With this sample size, we would be able to detect Cohen's *d* effect sizes of roughly .42 (at $p < .05$ with 80% power; G*Power; Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007).

2.2 Results

An Univariate Analysis of Variance (ANOVA) showed a significant effect of condition on perceptions of scientific quality $F(2, 204) = 4.39$, $p = .014$, $\eta_p^2 = .041$. Tukey's post-hoc analysis revealed that participants evaluated the anti-free will research as significantly less valid in the morally bad condition ($M = 3.18$, $SD = 1.20$; $n = 69$) than the neutral condition ($M = 3.77$,

SD = 1.17; *n* = 68), *p* =.009, Cohen's *d* = .50, consistent with prior research (Clark et al., 2014). As predicted, evaluations of the anti-free will research in the morally good condition (*M* = 3.48, *SD* = 1.10; *n* = 70) fell between the morally bad (*p* = .301; Cohen's *d* = .26) and morally neutral (*p* = .294; Cohen's *d* = .26) conditions, but these differences failed to reach statistical significance.[2]

Because these results were statistically inconclusive, we also conducted a linear trend analysis, which revealed a significant linear trend demonstrating a decrease in motivation to believe in free will from the morally bad to morally good to morally neutral conditions, *F*(1, 204) = 8.79, *p* = .003. Though note that the ANOVA in the preceding paragraph was planned, whereas this analysis was not.



[2] Repeating these analyses without participant exclusions revealed a marginal effect, *F*(2, 228) = 2.85, *p* = .059. Tukey's post-hoc analysis showed significant differences (*p* = .048; Cohen's *d* = .37) between the morally bad (*M* = 3.21, *SD* = 1.13) and the morally neutral conditions (*M* = 3.64, *SD* = 1.19). No other significant differences were found.

Figure 1. Scientific quality ratings by condition in Study 1a (scale range: 1-7).

2.3 Discussion

Consistent with expectations, participants were most critical of the anti-free will passage in the morally bad condition, followed by the morally good condition, followed by the neutral condition. However, the morally good condition did not significantly differ from either. A possible explanation for the weak findings was that our manipulation was too subtle. For this reason, we followed up with Study 1b, in which participants read about concrete instances of morally good and morally bad behavior in order to elicit stronger desires to blame and praise. We then conducted mini-meta-analyses of the results of these two initial studies to increase confidence in this pattern of findings.

3. Study 1b

In Study 1b, participants read what was ostensibly a brief news report about a morally bad, morally good, or morally neutral behavior, and then evaluated how freely performed the behavior was and reported their desire to blame or praise the actor. Participants also completed a general measure of free will belief. For both free will attributions and free will beliefs, we predicted that free will judgments would be highest in the morally bad condition, followed by the morally good condition, followed by the neutral condition. Furthermore, we expected desires to blame to mediate the difference between the morally bad and neutral conditions on both free will attributions and free will beliefs, and we expected desires to praise to mediate the difference between the morally good and neutral conditions on both free will attributions and free will beliefs.

3.1 Method

Participants (175 undergraduates; 125 Female; M_{age} = 20.65) were randomly assigned to read one of three ostensible news stories about 1) a thief stealing expensive lasers from a pediatric cancer hospital (morally bad condition), 2) a hospital administrator purchasing expensive lasers for a pediatric cancer hospital (neutral condition), or 3) a contributor donating expensive lasers to a pediatric cancer hospital (morally good condition). The target sample size was 180 (~60/condition, based on the experimenter's experience conducting similar research); 5 participants who signed up for the study failed to complete the experimental procedures, resulting in 175.[3] With this sample size, we would be able to detect Cohen's *d* effect sizes of roughly .47 (at $p < .05$ with 80% power; G*Power; Faul et al., 2009; 2007).

Participants then rated the extent to which the actor behaved of his own free will on three items: whether the action was freely chosen, whether the actor could have made other choices, and whether the actor exercised his own free will in choosing to perform the action on 7-point scales from *Not at all* to *Very much so,* $\propto$ = .92. Participants in the morally good and neutral conditions then reported the extent to which the actor was morally praiseworthy, and participants in the morally bad and neutral conditions reported the extent to which the actor was morally blameworthy on 7-point scales from *Not at all* to *Very much so*. Last, participants reported their free will beliefs on the FAD-Plus (Paulhus & Carey, 2011; $\propto$ = .77).[4] Participants also reported demographic information.

3.2 Results

3.2.1 Free will attributions. An ANOVA revealed a significant effect of condition on free will attributions, $F(2, 170) = 24.60$, $p < .001$, $\eta_p^2 = .224$ (see Figure 2). As hypothesized,

[3] Cory Clark was the experimenter for this study. Clark et al. (2017) aimed for approximately 60 participants per condition in each of their three experimental studies (Studies 3-5).
[4] Participants also completed the scientific and fatalistic determinism subscales of the FAD-Plus, but these were not analyzed.

Tukey's post-hoc tests revealed that participants attributed more free will to the morally good actor (n = 49; M = 5.71, SD = 1.15) than the morally neutral actor (n = 54; M = 4.56, SD = 1.07), p<.001, Cohen's d = 1.04. Replicating prior work, participants also attributed more free will to the morally bad actor (n = 70; M = 6.05, SD = 1.32) than the morally neutral one, p<.001, Cohen's d = 1.24 (Clark et al., 2014). As in Study 1a, participants attributed more free will to the bad actor than to the good actor with a small to medium effect size, Cohen's d = .27, but this difference failed to reach statistical significance, p = .297.[5]

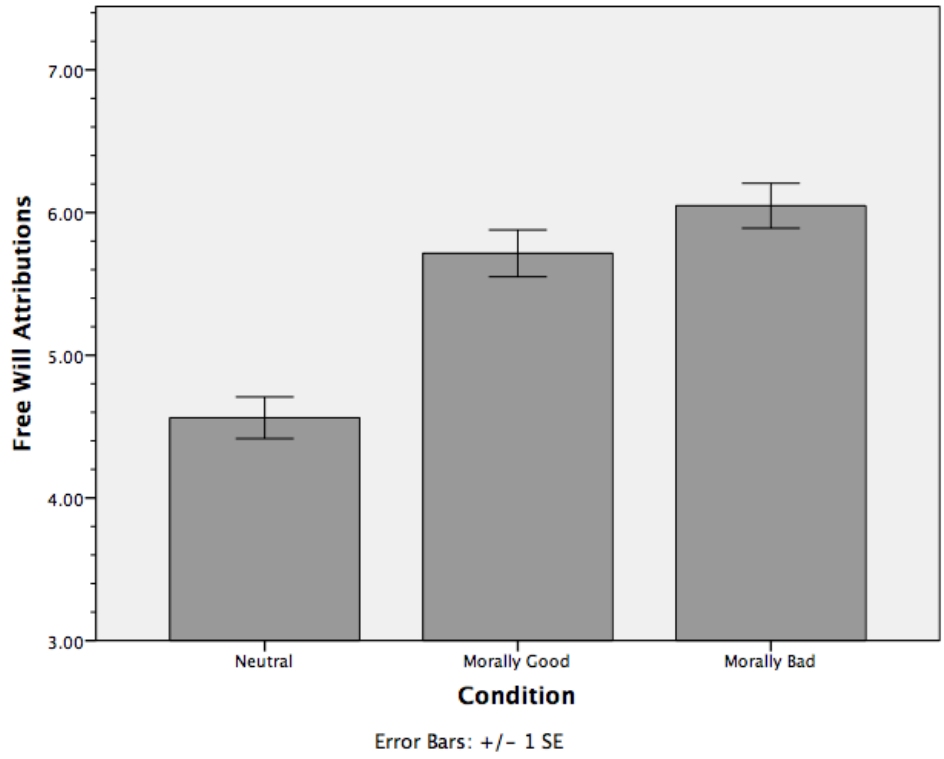


Figure 2. *Attributions of free will by condition in Study 1b* (scale range: 1-7).

3.2.2 Free will beliefs. There was a marginal effect of condition on general free will beliefs, $F(2, 165) = 2.81$, p = .063, η_p^2 = .033 (see Figure 3). Mirroring the pattern for free will attributions, free will beliefs were higher in the morally good condition (n = 47; M = 3.68, SD =

[5] Please see the supplementary file under the heading Unreported Studies X and Y, which describes two similar studies that replicated this pattern of results.

.51) than the neutral condition (n = 51; M = 3.49, SD = .72), though this difference failed to reach statistical significance, p = .272, Cohen's d = .30. Free will beliefs were marginally higher in the morally bad condition (n = 70; M = 3.75, SD = .56) than the neutral condition, p = .053, Cohen's d = .40, whereas the morally bad and morally good conditions did not differ, p = .805, Cohen's d = .13.

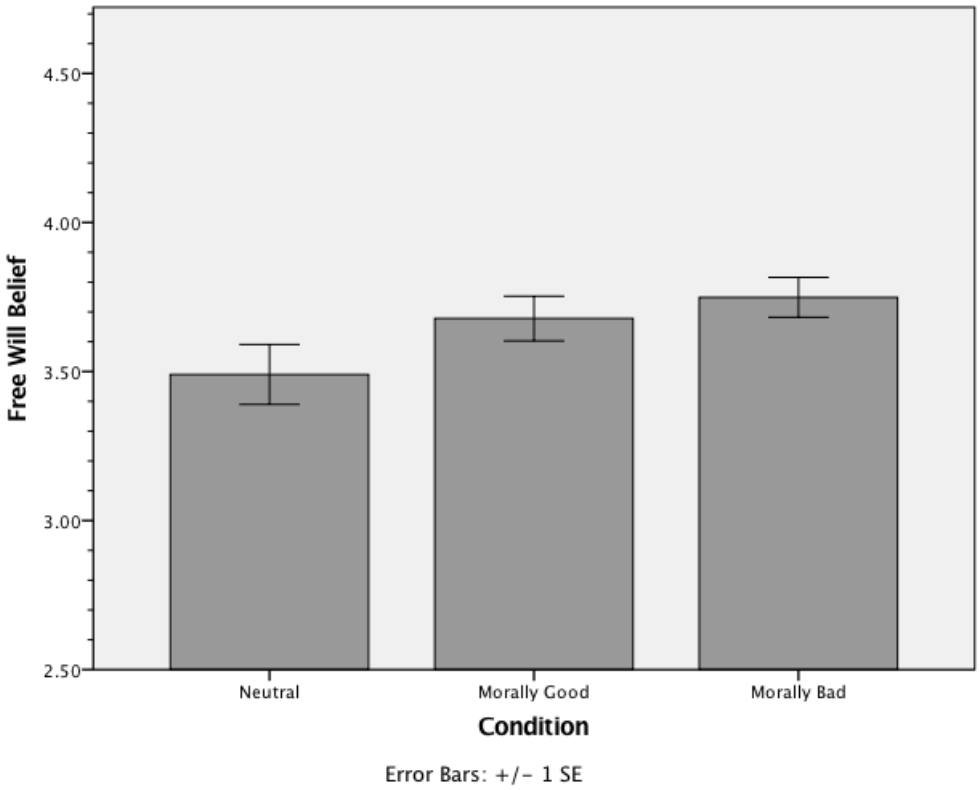


Figure 3. *Free will beliefs by condition in Study 1b* (scale range: 1-5).

3.2.3 Praise and blame. Independent samples t-tests revealed that participants felt the morally good actor was more morally praiseworthy (n = 49; M = 6.00, SD = 1.21) than the morally neutral actor (n = 52; M = 5.10, SD = 1.26), $t(99) = 3.68$, p <.001, Cohen's d = .73, and that the morally bad actor was more morally blameworthy (n = 72; M = 5.69, SD = 1.89) than the morally neutral actor (n = 54; M = 3.24, SD = 1.59), $t(124) = 7.71$, p <.001, Cohen's d = 1.40.

A bootstrap mediation analysis (10,000 resamples; PROCESS macro [Hayes, 2013]) revealed that perceptions that the morally good actor was more deserving of moral praise partially accounted for higher attributions of free will to the morally good actor, 95% CI [.109, .682]. This model is depicted in Figure 4. A similar but weaker pattern emerged for general free will beliefs, such that controlling for perceived praiseworthiness reduced the effect size of the neutral vs. good condition on free will belief, from semipartial $r = .15$, $p = .156$, to semipartial $r = .06$, $p = .523$, though the mediation was not statistically significant, 95% CI [-.024, .362]. This is likely due to the relatively small difference in free will beliefs between the neutral and morally good conditions.

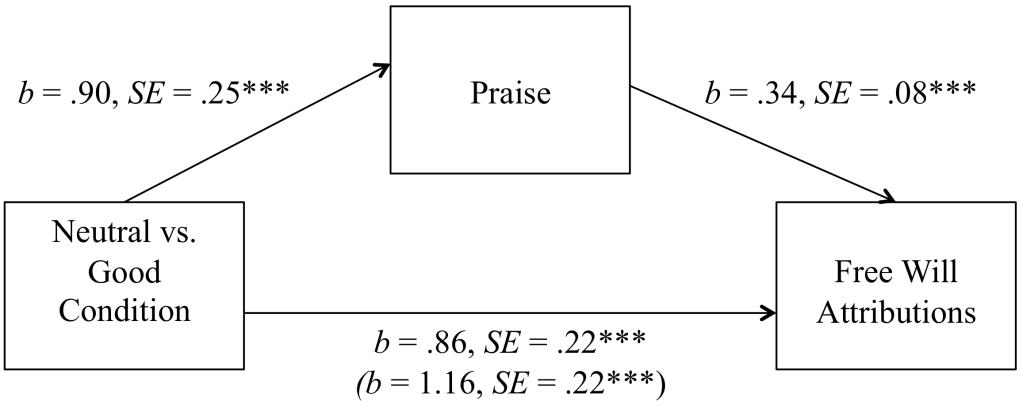


Figure 4. Influence of condition (Neutral: 0, Good: 1) on free will attributions mediated by praise in Study 1b. Values in parenthesis are the unstandardized regression coefficient and standard error prior to controlling for praise. ***p<.001.

Replicating Clark and colleagues (2014), perceptions that the morally bad actor was more deserving of moral blame partially accounted for higher attributions of free will to the morally bad actor, 95% CI [.484, 1.429],[6] and fully mediated the influence of the morally bad condition on general free will beliefs, 95% CI [.020, .345].

[6] Please see the supplementary file under the heading Unreported Studies X and Y for results of two similar studies, in which one study replicated this mediation pattern and one did not.

3.3 Discussion

With closely matched behaviors, participants attributed more free will to the morally good actor than the morally neutral actor, and this relationship was mediated by stronger desires to praise the morally good actor. Though a similar pattern emerged for general free will beliefs, the results were not statistically significant. As expected, these relationships were generally stronger in the morally bad conditions such that participants attributed the most free will to the morally bad actor and believed marginally more in free will in the morally bad condition. Further, greater desires to blame mediated the influence of the morally bad condition on both free will attributions and free will beliefs. It is clear that harmful actions that elicit desires to blame and punish have a very strong influence on free will judgments, but the present results indicate that morally good actions that elicit desires to praise have a similar albeit weaker effect.

Undoubtedly, people ascribe more blame and praise to an actor who had more freedom and control over causing a particular harmful or helpful outcome (e.g., consider the distinctions between murder, voluntary manslaughter, and involuntary manslaughter). The present results suggest that the reverse causal direction may also occur: people may increase perceptions that a behavior was performed freely when desiring to blame or praise. Of course, it is challenging to make causal conclusions from mediation models; often, both causal patterns (IV to mediator to DV and IV to DV to mediator) are statistically significant. This was the case in Study 1b as well: all significant mediation models were also significant if the mediator and DV were reversed. Nonetheless, our proposed causal order might make more sense a priori. As our IVs were intended to manipulate the mediator directly, it seems less likely that participants would only desire to blame or praise harmful and helpful actions *because* they were perceived as more freely performed. Independent of increased perceptions of freedom, it is likely that people would

desire to praise morally good actions simply because they are morally good and blame morally bad actions simply because they are morally bad. Thus, it seems likely that there is bidirectional causality such that high perceptions of freedom increase blame and praise, but desires to blame and praise can also increase perceptions of freedom.

4. Mini Metas

Because some of the results of Studies 1a and 1b were statistically inconclusive, we conducted three mini meta-analyses of the effect sizes of the differences in free will judgments between neutral and good, neutral and bad, and good and bad. In recent years, psychological science has encouraged greater emphasis on meta-analyses of overall patterns of findings across studies rather than on the statistical significance of individual studies (e.g., Maner, 2014). A meta-analysis of the results of studies testing similar questions provides better evidence of whether a given effect is real than the statistical significance of individual tests (e.g., Braver, Thoemmes, Rosenthal, 2014). We employed many of the suggested procedures outlined by Goh, Hall, and Rosenthal (2016) to conduct our three mini metas. Given that a large body of work has demonstrated that people attribute more responsibility to morally bad than morally good actions and neutral actions, we were primarily interested in meta-analyzing whether people attributed significantly more free will to morally good actors than morally neutral ones. Because we are specifically interested in interpreting the statistical significance of Studies 1a and 1b (particularly for the good vs. neutral comparison), we interpret fixed effects (Goh et al., 2016). Random effects are likely too conservative to detect significant effects with such a limited number of studies, nonetheless, we also report random effects as suggested by Goh and colleagues (2016; see also Hedges & Vevea, 1998).

We included one effect size for each study (i.e., the effect sizes for free will attributions and free will beliefs were averaged in Study 1b; Card, 2012), thus two effect sizes were included in each mini meta (the minimum necessary for a meta-analysis; Goh et al., 2016; Valentine, Pigott, & Rothstein, 2010). We computed *r* effect sizes from the *M*s, *SD*s, and *n*s. For fixed effects, all *r*s were Fisher's *z* transformed to r_z which were then weighted and averaged using the following formula: Weighted $\bar{r_z} = \Sigma ([N-3]r_z) / \Sigma (N-3)$. The weighted $\bar{r_z}$ s were then converted back to Pearson correlations for presentation. To determine statistical significance, we utilized the Stouffer's *Z* test, in which the *p* values for each *r* effect size were converted to *Z*s. The *Z*s were then combined using the following formula: $Z_{combined} = \Sigma Z / \text{sqrt}(k)$, and then converted back to *p*s for presentation. For random effects, we conducted one-sample *t*-tests of the effect sizes.[7] All of these procedures are more thoroughly described in Goh et al. (2016).

As can be seen in summary Table 1, consistent with our main hypothesis, free will judgments were significantly higher in the good than neutral conditions with a small to medium effect size, $r = .21$, $p = .0007$ (random effects: $r = .23$, $p = .254$). Consistent with Clark and colleagues (2014), free will judgments were significantly higher in the bad than neutral conditions, $r = .30$, $p < .0001$ (random effects: $r = .31$, $p = .134$). There was also a marginal effect such that free will judgments were higher in the morally bad than morally good conditions, $r = .12$, $p = .064$ (random effects: $r = .12$, $p = .083$).[8,9]

[7] These data and the syntax are publically available.

[8] If we add the effect sizes for the good vs. bad comparisons in the upcoming Studies 2a and 2b, this difference is significant, $r = .14$, $p < .0001$ (random effects: $r = .14$, $p = .006$). This relationship is also consistent with the large body of similar research showing higher attributions of responsibility for morally bad than morally good actions. Thus we feel there is little reason to doubt that this effect is real, even if it is somewhat small here.

[9] If we also add the effect sizes for two unreported studies that were similar in design to Study 1b (see Unreported Studies X and Y and Table 2S in Supplementary File), all conditions significantly differ, *p*s < .0001 (random effects: *p*s < .020).

Table 1. *Meta-analyzed Effect Sizes for Good vs. Bad, Good vs. Neutral, and Bad vs. Neutral*

	Free Will Measure/s		**Comparison**		
			Good vs. Bad	Good vs. Neutral	Bad vs. Neutral
Study 1a	Evaluation of Anti-Free Will Research (reversed)	*r*	0.13	0.13	0.24
		n	139	138	137
		p	0.127	0.129	0.0047
Study 1b	Attributions and Beliefs (combined)	*r*	0.10	0.32	0.37
		n	119	103	124
		p	0.279	0.001	<0.0001
Unweighted (random)		***r***	**0.12**	**0.23**	**0.31**
		p	**0.083**	**0.254**	**0.134**
Weighted (fixed)		***r***	**0.12**	**0.21**	**0.30**
		p	**0.064**	**0.0007**	**<.0001**

5. Study 2a

The mini meta of Studies 1a and 1b supported our first hypothesis that free will judgments would be higher for morally positive actions than morally neutral ones, and (to some extent) that free will judgments would be lower for morally positive actions than for morally bad ones (see Footnote 8 for mini meta-analysis results including upcoming Study 2a and 2b data, which confirm this relationship). In Studies 2a and 2b, we moved on to test our second hypothesis that free will judgments for immoral actions would be predominantly sensitive to affective punitive motives, whereas free will judgments for morally positive actions would be sensitive to other relevant information such as the utility of reward. Research has shown that the motives underlying punishment appear to be predominantly retributive and that the actual utility of punishment has little influence on punishment decisions (e.g., Carlsmith et al., 2002). When people are exposed to others' morally bad actions, they experience strong impulses to punish the harmdoer, and they are relatively insensitive to considerations such as whether the punishment

would deter bad behavior. Therefore, we expected that free will judgments for immoral actions would be sensitive to retributive concerns only.

In contrast, positive emotional reactions to morally good actions may allow individuals to more critically evaluate various aspects of the act itself, including the likely utility of the reward. Relative to negative emotions, positive emotions broaden the scope of awareness and allow people to consider more relevant information in their decision-making. Therefore, when deciding whether a morally good act is deserving of praise, people might consider the likely utility of the reward, and other features of the act in question. We expected that free will judgments for morally positive actions would be more contextually sensitive than those for immoral actions, and in particular, more sensitive to utilitarian concerns.

Methods for Studies 2a and 2b were adapted from studies by Carlsmith and colleagues (2002), in which the deservingness of punishment (retributive motive) was manipulated by varying the severity of the immoral act, and utility (deterrent motive) was manipulated by varying the publicity of the punishment. We applied this paradigm to the present work by having participants read about either a morally good or morally bad act and manipulating the deservingness and utility of the punishment or reward. As in Carlsmith and colleagues' (2002) studies, deservingness was manipulated by varying the magnitude of the action, and utility was manipulated by specifying whether the reward or punishment would become public knowledge with the explicit purpose of either deterring others from doing the same (in the morally bad condition) or encouraging others to do the same (in the morally good condition).

Participants then recommended a punishment or reward for the actor, rated the magnitude of the action, rated the likely utility of the punishment or reward, rated how freely performed the action was, and reported their general free will belief. We expected that in the morally bad

condition, people would punish more severely, attribute more free will, and believe more in free will in the high deservingness condition than the low deservingness condition, but that these judgments would not differ between the high and low utility conditions. In the morally good condition, we expected that reward, free will attributions, and free will beliefs would be higher in the high deservingness condition than the low deservingness condition, but also in the high utility condition than the low utility condition. These results would suggest that praise motives and ascriptions of free will for morally positive behaviors are sensitive to both deservingness and utility.

For exploratory purposes, participants also reported their emotional reaction, rated how normative the behavior was, and rated how much willpower was exerted. If free will judgments for immoral actions are primarily driven by affective punitive responses, these judgments should be predicted most strongly by participants' emotional reactions. If free will judgments for morally positive actions are more carefully considered and context-dependent, they might be better predicted by ratings of normativity and required willpower.

5.1 Method

In a 2 (Deservingness: High vs. Low) × 2 (Utility: High vs. Low) × 2 (Moral Valence: Bad vs. Good) between-subjects design, participants (451 undergraduates; 285 female; M_{age} = 19.27) were randomly assigned to read one of eight short stories about an ostensible recent event, involving John, a district manager at a large department store, who performed either a morally admirable or morally deplorable action. The target sample size was 480 (again ~60/condition)[10] or however many participants could be recruited by the end of the semester. The end of the semester came first, resulting in 451 participants. Participants were randomly assigned to read

[10] The target numbers of participants per condition were the same in Studies 1b and 2a because the same experimenter (Cory Clark) performed these studies.

that John had either falsely reported earnings to steal from the company (morally bad condition) or sacrificed his own bonus to implement an employee incentive program that increased profits for the company (morally good condition). The amount of money stolen or earned was also manipulated: $9.2 million in the high deservingness conditions vs. $24,000 in the low deservingness conditions. Last, the utility of John's punishment or reward was manipulated by indicating whether the punishment or reward would become public knowledge in order to deter others (in the morally bad condition) or encourage others (in the morally good condition).

Participants then rated John's free will on the same three free will attribution items from Study 1b, on 9-point scales. Participants in the morally good condition then reported how much John should be rewarded and participants in the morally bad condition reported how much John should be punished on two questions. The first question asked how severely John should be punished or how highly John should be rewarded on a 7-point scale from *Not at all* to *Extremely*. They were then given concrete punishment or reward options. Participants in the morally bad condition were asked what an appropriate prison sentence for John would be with 14 options increasing in severity from *no prison* to *50 years to life*. Participants in the morally good condition were asked what an appropriate reward bonus for John would be with 15 options increasing in amount from *no bonus* to *more than $1,000,000*. The two punishment questions and the two reward questions were converted to *Z*-scores and combined into individual punishment and reward indices.

Participants then evaluated the magnitude of the action by reporting either how serious John's crime was or how generous John's actions were, which also served as a deservingness manipulation check. Participants also rated the utility of the response by indicating how likely it is that John's punishment would deter others from doing the same thing or that John's reward

would encourage others to do the same thing, which also served as a utility manipulation check. To gauge emotional reactions, participants were asked how morally outraged or morally uplifted they were by John's behavior. Action magnitude, response utility, and emotional reactions all were reported on 7-point scales from *Not at all* to *Extremely*.

Participants then rated how normative the behavior was by indicating how likely it is that someone else in John's position would have done the same thing rated on a 7-point scale from *Not at all likely* to *Extremely likely* and how much John used his willpower rated on a 7-point scale from *Not at all* to *Completely*. Last, participants completed the free will subscale of the FAD-Plus (Paulhus & Carey, 2011), reported demographic information, and were debriefed.

5.2 Results

5.2.1 Action magnitude manipulation check. A 2 (deservingness) × 2 (utility) × 2 (moral valence) ANOVA revealed a main effect of the deservingness manipulation on action magnitude ratings, $F(1, 443) = 13.06$, $p < .001$, $\eta^2_p = .029$, such that participants in the high deservingness conditions rated John's actions as more serious/generous ($M = 5.97$, $SD = 1.09$; $n = 228$) than participants in the low deservingness conditions ($M = 5.64$, $SD = 1.20$; $n = 223$).[11]

5.2.2 Response utility manipulation check. A 2×2×2 ANOVA revealed a main effect of the utility manipulation, $F(1, 443) = 9.86$, $p = .002$, $\eta^2_p = .022$, such that participants in the high utility conditions reported that John's punishment/reward would be more likely to discourage/encourage others ($M = 4.89$, $SD = 1.41$; $n = 224$) than participants in the low utility conditions ($M = 4.42$, $SD = 1.82$; $n = 227$).

[11] For space purposes, only significant or marginal effects that are relevant to the present hypothesis are reported in text.

5.2.3 Punishment. A 2 (deservingness) × 2 (utility) ANOVA on the punishment index revealed just one main effect for deservingness, $F(1, 219) = 16.07$, $p < .001$, $\eta^2_p = .068$, such that participants in the high deservingness condition recommended harsher punishment ($M = .22$, $SD = .91$; $n = 116$) than those in the low deservingness condition ($M = -.24$, $SD = .77$; $n = 107$).

5.2.4 Reward. A 2 (deservingness) × 2 (utility) ANOVA on the reward index revealed a main effect for deservingness, $F(1, 222) = 4.51$, $p = .035$, $\eta^2_p = .020$. Participants in the high deservingness condition recommended higher reward ($M = .12$, $SD = .95$) than those in the low deservingness condition ($M = -.12$, $SD = .81$). There was no main effect for the utility condition, but there was an interaction between the utility and deservingness manipulations, $F(1, 222) = 4.75$, $p = .030$, $\eta^2_p = .021$. In the low utility condition, deservingness of the reward had no influence on how highly participants rewarded (high deservingness $M = .017$, $SE = .12$; $n = 56$; low deservingness $M = .023$, $SE = .12$; $n = 58$), $F(1, 222) = .002$, $p = .968$, $\eta^2_p = .000$. In the high utility condition, participants in the high deservingness condition recommended higher rewards ($M = .24$, $SE = .12$; $n = 54$) than participants in the low deservingness condition ($M = -.27$, $SE = .12$; $n = 58$), $F(1, 222) = 9.17$, $p = .003$, $\eta^2_p = .040$. See Figure 5 for graphs of the interactions between the deservingness and utility manipulations on desires to punish and reward in both Studies 2a and 2b.

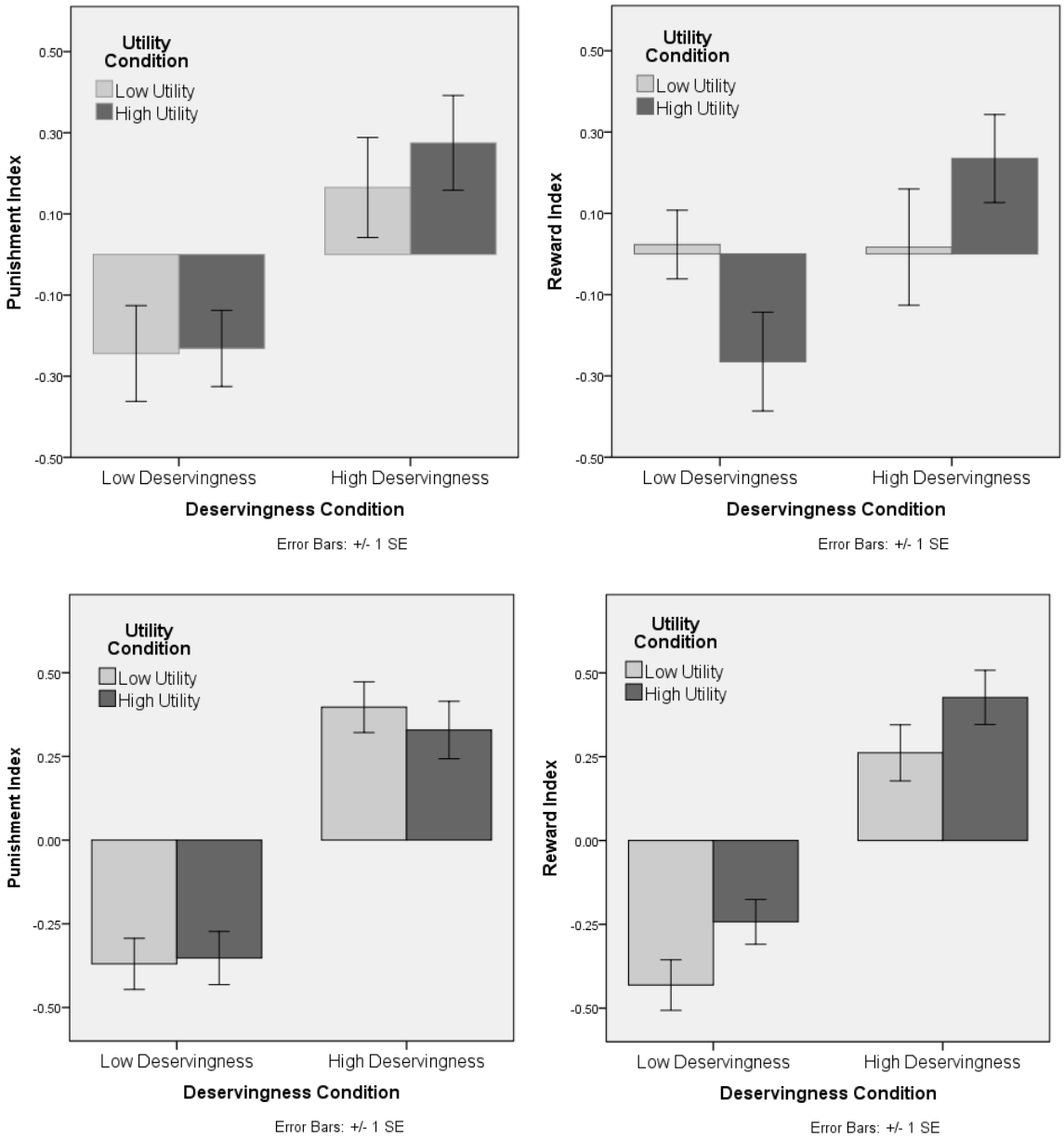


Figure 5. Interactions between the deservingness and utility manipulations on desires to punish (left column) and reward (right column) in Study 2a (top row) and Study 2b (bottom row).

5.2.5 Free will judgments. For free will attributions, there was only a main effect of the moral valence condition, $F(1, 443) = 30.37$, $p < .001$, $\eta^2_p = .064$, such that participants attributed more free will to the morally bad actor ($M = 8.52$, $SD = .93$; $n = 223$) than to the morally good

actor (M = 7.86, SD = 1.53; n = 228). The same result was found for general free will belief. Participants in the morally bad condition believed more in free will (M = 4.00, SD = .64; n = 222) than participants in the morally good condition (M = 3.84, SD = .70; n = 228), $F(1, 442)$ = 6.74, p = .010, η^2_p = .015. Contrary to expectations, there were no interactions with deservingness or utility on either free will attributions or beliefs.

5.2.6 Exploratory analyses. Study 2a replicated the results of the mini meta and prior work demonstrating that people ascribe more free will to morally bad actions than morally good ones, and the deservingness and utility manipulations worked as expected. Also as expected, punishment was only sensitive to the deservingness manipulation whereas reward was sensitive to both deservingness and utility. However, against our hypothesis, no interactions were found between the deservingness or utility manipulations with the moral valence condition on free will attributions or beliefs.

We hypothesized that free will judgments for morally bad actions would be driven primarily by affective retributive concerns (and not utility concerns), but that free will judgments for morally good actions would be sensitive to utility concerns. Despite the failure of our manipulations to demonstrate such findings, we included a variety of other evaluation variables that could help test this hypothesis. For this reason, we conducted additional exploratory analyses to test whether free will judgments for morally bad actions were primarily predicted by participants' ratings of the magnitude of the immoral act, emotional responses, and desires to punish (outcomes relevant to affect and retribution), whereas free will judgments for morally good actions might be more sensitive to less emotional features such as whether the reward would be likely to encourage good behavior, how much willpower the action required, and the degree of counternormativity of the action.

We analyzed the interactions between the moral valence condition and each of the six following evaluation variables: action magnitude ratings, response utility ratings, emotional reaction, normativity ratings, willpower ratings, and punish/reward response on both free will attributions and free will beliefs. Six separate regressions were performed for both free will attributions and free will beliefs. Each included the moral valence condition, one of the six evaluation variables, and the interaction between the moral valence condition and that variable. As can be seen in Tables 2 (free will attributions) and 3 (free will beliefs), there were significant main effects for all six evaluation variables on both free will attributions and beliefs (with the exception of normativity ratings on free will beliefs), such that higher free will judgments were predicted by higher ratings of action magnitude, higher ratings that the punishment or reward would have utility, stronger emotional reactions, lower ratings of normativity, higher ratings of willpower, and stronger desires to punish or reward. Many of these were qualified by interactions with the moral valence condition, which will be described below.

5.2.6.1 Action magnitude. Higher ratings of action magnitude predicted higher attributions of free will in both the morally good and morally bad conditions. Simple slopes within the morally good and morally bad conditions demonstrated that higher action magnitude ratings were stronger predictors of higher free will attributions in the morally good condition (b =.75), $t = 10.54$, $p < .001$, than in the morally bad condition (b =.21), $t = 3.00$, $p = .003$. This interaction did not emerge for free will beliefs. Rather, action magnitude ratings predicted higher free will beliefs in both the morally good and bad conditions to equal degrees.

5.2.6.2 Response utility. There was no interaction between the moral valence condition and response utility on free will attributions or free will beliefs. Higher utility of the punishment and reward both predicted higher free will attributions and beliefs.

5.2.6.3 Emotional reaction. In the morally bad condition, greater moral outrage significantly predicted higher free will attributions (b =.54), t = 9.84, p < .001. However, this effect was larger in the morally good condition such that feeling morally uplifted more strongly predicted higher free will attributions (b =.75), t = 13.66, p < .001. There was no interaction for free will beliefs. A stronger emotional response predicted higher free will beliefs in both the morally good and bad conditions to equal degrees.

5.2.6.4 Normative. Participant ratings of the likelihood that other people in the same position would have done the same thing did not predict free will attributions to the morally bad actor (b = -.01), t = -0.18, p = .855. In contrast, participants attributed more free will to the morally good actor as their perceived likelihood that others in the same position would have done the same thing decreased (b =-.21), t = -3.26, p = .001. Normativity ratings did not predict free will beliefs in either condition.

5.2.6.5 Willpower. In the morally bad condition, there was only a marginal relationship between willpower ratings and free will attributions (b =.06), t = 1.77, p = .077, and no relationship between willpower ratings and free will beliefs (b =.02), t = 1.36, p = .175. In the morally good condition, participants who felt that the actor exerted more willpower attributed him more free will (b =.54), t = 9.88, p < .001, and believed more in free will generally (b =.09), t = 2.81, p = .005.

5.2.6.6 Punish/reward response. There was no interaction between the condition on the extent to which desires to reward or punish increased free will attributions and beliefs. Stronger desires to reward predicted higher free will attributions and beliefs in the morally good condition and stronger desires to punish predicted higher free will attributions and beliefs in the morally bad condition.

Table 2
Free Will Attributions Regressed on Moral Valence Condition, Evaluation Variables, and Interactions in Study 2a (Exploratory Analyses)

	F	***R^2***	***b***	***SE***	***t***	***p***	***95% CIs***
Free Will Attributions (Model)	65.97	.51				<.001	
Moral valence condition			1.04	.11	9.16	<.001	.82, 1.26
Action magnitude			.75	.07	10.30	<.001	.60, .89
Interaction			**-.53**	**.10**	**-5.42**	**<.001**	**-.73, -.34**
Free Will Attributions (Model)	11.92	.07				<.001	
Moral valence condition			.69	.12	5.77	<.001	.38, .66
Response utility			.10	.05	2.06	.040	.04, .17
Interaction			**-.06**	**.07**	**-0.78**	**.435**	**-.20, .09**
Free Will Attributions (Model)	21.94	.13				<.001	
Moral valence condition			.75	.12	6.41	<.001	.52, .98
Emotional reaction			.31	.06	5.41	<.001	.20, .43
Interaction			**-.21**	**.08**	**-2.65**	**.008**	**-.36, -.05**
Free Will Attributions (Model)	14.04	.09				<.001	
Moral valence condition			.75	.12	6.06	<.001	.50, .99
Normative			-.21	.06	-3.30	.001	-.33, -.08
Interaction			**.20**	**.09**	**2.27**	**.024**	**.03, .37**
Free Will Attributions (Model)	42.55	.22				<.001	
Moral valence condition			.99	.12	8.63	<.001	.77, 1.22
Willpower			.54	.06	9.39	<.001	.43, .66
Interaction			**-.49**	**.07**	**-7.29**	**<.001**	**-.62, -.35**
Free Will Attributions (Model)	19.41	.12				<.001	
Moral valence condition			.66	.12	5.68	<.001	.43, .89
Punish/reward response			.37	.09	4.01	<.001	.19, .56
Interaction			**-.08**	**.13**	**-.57**	**.571**	**-.34, .19**

Note. For ease of comparison, **bold** indicates interactions that are significant in Study 2b, which had a larger sample size.

Table 3
Free Will Beliefs Regressed on Moral Valence Condition, Evaluation Variables, and Interactions in Study 2a (Exploratory Analyses)

	F	***R2***	***b***	***SE***	***t***	***p***	**95% *CI*s**
Free Will Belief (Model)	9.73	.06				<.001	
Moral valence condition			.27	.07	4.11	<.001	.14, .40
Action magnitude			.13	.04	3.05	.002	.05, .21
Interaction			**.01**	**.06**	**0.18**	**.856**	**-.10, .12**
Free Will Belief (Model)	8.80	.06				<.001	
Moral valence condition			.20	.06	3.19	.002	.08, .32
Response utility			.08	.03	3.20	.001	.03, .13
Interaction			.00	.04	0.07	.948	-.07, .08
Free Will Belief (Model)	10.53	.07				<.001	
Moral valence condition			.21	.06	3.34	.001	.09, .33
Emotional reaction			.12	.03	3.89	<.001	.06, .18
Interaction			**-.03**	**.04**	**-0.77**	**.444**	**-.12, .05**
Free Will Belief (Model)	3.18	.02				.024	
Moral valence condition			.13	.07	2.03	.043	.00, .26
Normative			.04	.03	1.25	.212	-.02, .11
Interaction			-.01	.05	-0.15	.881	-.10, .08
Free Will Belief (Model)	5.11	.03				.002	
Moral valence condition			.23	.07	3.42	.001	.10, .36
Willpower			.09	.03	2.67	.008	.02, .15
Interaction			**-.07**	**.04**	**-1.73**	**.084**	**-.14, .01**
Free Will Belief (Model)	7.81	.04				<.001	
Moral valence condition			.17	.06	2.64	.008	.04, .29
Punish/reward response			.13	.05	2.69	.007	.04, .23
Interaction			.02	.07	0.27	.785	-.12, .16

Note. For ease of comparison, **bold** indicates interactions that are significant in Study 2b, which had a larger sample size.

5.3 Discussion

Study 2a explored how deservingness and utility differentially predicted punitive and reward responses as well as free will judgments for morally good and bad behaviors. Replicating the results of the mini meta and prior work, participants attributed more free will to the morally bad actor than the morally good actor. Furthermore, as hypothesized, participants were primarily

concerned with how deserving the immoral actor was when determining punishment. Participants advocated harsher punishments for more severe misdeeds, regardless of whether anyone else would find out, and thus regardless of whether the punishment would be likely to deter others from committing similar crimes. When determining rewards, there was some evidence indicating that participants were concerned with both deservingness and the likelihood that the reward would improve others' moral behavior. Participants only advocated greater rewards for more virtuous actions when the rewards would be publicly known. However, these differences did not manifest in free will judgments as we had hypothesized.

Nonetheless, the exploratory analyses revealed a number of useful findings. Specifically, free will judgments for morally bad actions were predominantly predicted by responses related to retributive feelings (action magnitude ratings, emotional response, and desires to punish), but not the less emotional responses (willpower and normativity ratings). The main exception was perceived utility, which predicted higher free will attributions and beliefs for both morally good and bad actions. Free will judgments for morally good actions were predicted by all of these, including willpower and normativity.

Though these results seem generally consistent with our hypothesis that free will judgments for morally good actions are more contextually sensitive than free will judgments for morally bad acts, whereas free will judgments for morally bad acts are predominantly driven by affective punitive motives, they were not predicted a priori. To increase confidence that these effects were real, we conducted a near replication of Study 2a in order to confirm these results. Furthermore, many of these effects emerged only for free will attributions and not free will beliefs. We may have failed to find similar effects for free will beliefs because they were measured at the very end of the study, or because free will beliefs are generally more stable and

harder to manipulate. Nonetheless, we expected similar patterns of results for general free will beliefs as for free will attributions, even if the effect sizes would be smaller. For this reason, a larger sample size was collected for Study 2b.

6. Study 2b

Study 2b was designed to confirm and replicate Study 2a. Because some of the effects that emerged for free will attributions did not emerge for free will beliefs, we recruited a larger sample size to increase our ability to detect any similar effects on general free will beliefs and any smaller influences on free will attributions.

6.1 Method

Study 2b (n = 784 MTurk participants; 475 female; M_{age} = 34.66) was a near identical replication of Study 2a except the punish/response, action magnitude rating, utility rating, and emotional response came before the free will attributions, which included one additional item: "Relative to typical, every day sorts of actions, to what extent were John's actions more or less free?" rated on a 9-point scale from *Much less free* to *Much more free*, α = .74. Sample size was determined by the maximum number of participants that $650 could pay for at $0.65/participant including MTurk and TurkPrime fees.

6.2 Results

6.2.1 Action magnitude manipulation check. A 2×2×2 ANOVA replicated the main effect of the deservingness manipulation, $F(1, 776) = 48.66$, $p < .001$, $\eta^2_p = .059$, such that participants in the high deservingness conditions rated John's actions as more serious/generous (M = 6.28, SD = .95; n = 389) than participants in the low deservingness conditions (M = 5.79, SD = 1.16; n = 395).

6.2.2 Response utility manipulation check. A 2×2×2 ANOVA replicated the main effect of the utility manipulation, $F(1, 776) = 31.82$, $p < .001$, $\eta^2_p = .039$, such that participants in the high utility conditions reported that John's punishment/reward would be more likely to discourage/encourage others ($M = 4.99$, $SD = 1.42$; $n = 393$) than participants in the low utility conditions ($M = 4.35$, $SD = 1.80$; $n = 391$).

6.2.3 Punishment. As in Study 2a, a 2 (deservingness) × 2 (utility) ANOVA on the punishment index revealed just one main effect for deservingness, $F(1, 389) = 82.93$, $p < .001$, $\eta^2_p = .176$, such that participants in the high deservingness condition recommended harsher punishment ($M = .36$, $SD = .80$; $n = 196$) than those in the low deservingness condition ($M = -.36$, $SD = .77$; $n = 197$). There was again no main effect for the utility condition, $p = .749$.

6.2.4 Reward. As in Study 2a, a 2 (deservingness) × 2 (utility) ANOVA on the reward index revealed a main effect for deservingness, $F(1, 385) = 78.57$, $p < .001$, $\eta^2_p = .169$. Rather than an interaction between deservingness and utility (as in Study 2a), there was a main effect of utility, $F(1, 385) = 5.29$, $p = .022$, $\eta^2_p = .014$. Participants in the high deservingness condition recommended higher reward ($M = .34$, $SD = .81$; $n = 192$) than those in the low deservingness condition ($M = -.34$, $SD = .71$; $n = 197$). Participants in the high utility condition also recommended higher reward ($M = .09$, $SD = .80$; $n = 195$) than those in the low utility condition ($M = -.09$, $SD = .86$; $n = 194$).

6.2.5 Free will. As in Study 2a, for free will attributions, there was only a main effect of the moral valence condition, $F(1, 776) = 13.15$, $p < .001$, $\eta^2_p = .017$, such that participants attributed more free will to the morally bad actor ($M = 8.14$, $SD = 1.05$; $n = 393$) than to the morally good actor ($M = 7.85$, $SD = 1.21$; $n = 391$). The same result was found for general free will belief. Participants in the morally bad condition believed more in free will ($M = 4.06$, $SD =$

.61; n = 393) than participants in the morally good condition (M = 3.96, SD = .65; n = 391), F(1, 776) = 4.80, p = .029, η^2_p = .006.

6.2.6 Confirmatory analyses. We ran the same twelve regressions as in Study 2a. As can be seen in Tables 4 (free will attributions) and 5 (free will beliefs), all main effects for the six evaluation variables on free will attributions and free will beliefs replicated (again, with the exception of the normative evaluation on free will beliefs). Higher free will judgments were predicted by higher action magnitude, higher response utility, stronger emotional reaction, lower normativity, higher willpower, and stronger punish/reward response. Most of these were again qualified by interactions with the moral valence condition.

6.2.6.1 Action magnitude. As in Study 2a, higher action magnitude ratings predicted higher attributions of free will in both the morally bad and good conditions, but once again were stronger predictors in the morally good condition (b = .65), t = 10.29, p < .001, than in the morally bad condition (b = .16), t = 3.56, p < .001. Unlike Study 2a, this effect also emerged for free will beliefs. People increased their belief in free will as their perceived magnitude of the morally bad action increased (b = .11), t = 3.42, p = .001, but they did so to a stronger degree as their perceived magnitude of the morally good action increased (b =.25), t = 8.00, p < .001.

6.2.6.2 Response utility. As in Study 2a, response utility ratings predicted higher attributions of free will and higher free will beliefs. Unlike Study 2a, the interaction between the moral valence condition and response utility reached statistical significance for free will attributions. Perceptions that punishing the immoral actor would deter bad behavior did not predict free will attributions (b = -.001), t = -0.02, p = .985, but perceptions that rewarding the morally good actor would encourage good behavior did predict higher free will attributions to the

morally good actor (b = .12), t = 2.68, p = .007. Once again, the interaction did not emerge for free will beliefs.

6.2.6.3 Emotional reaction. As in Study 2a, greater moral outrage predicted higher free will attributions to the morally bad actor (b = .32), t = 10.12, p < .001, but feeling morally uplifted predicted higher free will attributions to the morally good actor to a significantly stronger degree (b = .56), t = 12.52, p < .001. Unlike Study 2a, this interaction also emerged for general free will beliefs (morally bad condition [b = .11], t = 2.45, p = .014; morally good condition [b = .17], t = 5.25, p < .001).

6.2.6.4 Normative. Unlike Study 2a, higher ratings of the likelihood that other people in John's position would have done the same thing predicted lower free will attributions to the morally bad actor (b = -.10), t = -3.19, p = .001. However, as in Study 2a, this relationship was stronger in the morally good condition, such that participants attributed less free will to John as their perceived likelihood that others in John's position would have done the same thing increased (b = -.21), t = -4.61, p < .001. As in Study 2a, normativity judgments did not predict free will beliefs in either condition.

6.2.6.5 Willpower. There were again no relationships between ratings of John's willpower and free will attributions (b = .00) or beliefs (b = .01) in the morally bad condition. In the morally good condition, participants who felt that John exerted more willpower attributed him more free will (b = .37), t = 8.27, p < .001, and believed more in free will generally (b = .15), t = 4.71, p < .001.

6.2.6.6 Punish/reward response. As in Study 2a, both higher punishment recommendations (b = .27), t = 4.27, p < .001, and higher reward recommendations (b = .47), t = 7.43, p < .001, predicted higher attributions of free will. Unlike Study 2a, the relationship was

significantly stronger in the morally good condition. Once again, this interaction did not emerge for free will beliefs.

Table 4
Free Will Attributions Regressed on Moral Valence Condition, Evaluation Variables, and Interactions in Study 2b (Confirmatory Analyses)

	F	***R^2***	***b***	***SE***	***t***	***p***	***95% CIs***
Free Will Attributions (Model)	47.44	.15				<.001	
Moral valence condition			.60	.08	7.42	<.001	.44, .76
Action magnitude			.65	.06	10.73	<.001	.53, .77
Interaction			-.49	.08	-6.44	<.001	-.64, -.34
Free Will Attributions (Model)	7.88	.03				<.001	
Moral valence condition			.33	.09	4.05	<.001	.17, .50
Response utility			.12	.04	3.19	.001	.05, .19
Interaction			-.12	.05	-2.42	.016	-.22, -.02
Free Will Attributions (Model)	51.41	.17				<.001	
Moral valence condition			.42	.08	5.56	<.001	.27, .57
Emotional Reaction			.46	.04	10.33	<.001	.37, .54
Interaction			-.24	.06	-4.10	<.001	-.37, -.13
Free Will Attributions (Model)	14.46	.05				<.001	
Moral valence condition			.37	.08	4.54	<.001	.21, .53
Normative			-.21	.04	-4.77	<.001	-.29, -.12
Interaction			.11	.06	1.83	.068	-.01, .22
Free Will Attributions (Model)	32.16	.09				<.001	
Moral valence condition			.59	.09	6.67	<.001	.41, .76
Willpower			.37	.05	8.15	<.001	.28, .45
Interaction			-.37	.05	-7.40	<.001	-.47, -.27
Free Will Attributions (Model)	27.44	.10				<.001	
Moral valence condition			.30	.08	3.80	<.001	.14, .45
Punish/reward response			.47	.07	7.08	<.001	.34, .60
Interaction			-.20	.09	-2.19	.029	-.38, -.02

Table 5
Free Will Beliefs Regressed on Moral Valence Condition, Evaluation Variables, and Interactions in Study 2b (Confirmatory Analyses)

	F	***R2***	***b***	***SE***	***t***	***p***	**95% *CI*s**
Free Will Belief (Model)	24.90	.09				<.001	
Moral valence condition			.23	.05	5.06	<.001	.14, .32
Action magnitude			.25	.04	7.27	<.001	.18, .32
Interaction			-.15	.04	-3.31	.001	-.23, -.06
Free Will Belief (Model)	9.90	.04				<.001	
Moral valence condition			.14	.05	3.16	.002	.05, .23
Response utility			.06	.02	3.15	.002	.02, .10
Interaction			.01	.03	0.27	.789	-.05, .06
Free Will Belief (Model)	23.63	.08				<.001	
Moral valence condition			.15	.04	3.43	.001	.06, .24
Emotional reaction			.17	.03	6.49	<.001	.12, .22
Interaction			-.06	.03	-1.72	.086	-.13, .01
Free Will Belief (Model)	2.22	.01				.008	
Moral valence condition			.11	.05	2.33	.020	.02, .20
Normative			.00	.02	-0.08	.936	-.05, .05
Interaction			-.03	.03	-0.83	.410	-.09, .04
Free Will Belief (Model)	13.73	.05				<.001	
Moral valence condition			.23	.05	4.66	<.001	.13, .33
Willpower			.15	.03	5.90	<.001	.10, .20
Interaction			-.14	.03	-4.74	<.001	-.19, -.08
Free Will Belief (Model)	11.34	.04				<.001	
Moral valence condition			.10	.04	2.24	.025	.01, .19
Punish/reward response			.10	.04	2.63	.009	.03, .17
Interaction			.07	.05	1.35	.178	-.03, .17

Table 6

Summary of the relationships between each Evaluation Variable and Free Will Attributions and Beliefs within each Moral Condition and indicators of which (if either) condition had a significantly stronger relationship in Studies 2A (exploratory) and 2B (confirmatory)

Study	Evaluation Variable	Outcome	Condition	Direction	Significant	Stronger Condition
Study 2a	Action Magnitude	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	
Study 2b	Action Magnitude	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	X
			Morally Bad	+	*	
Study 2a	Response Utility	Free Will Attributions	Morally Good	+	*	
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	
Study 2b	Response Utility	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	0		
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	
Study 2a	Emotional Reaction	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	
Study 2b	Emotional Reaction	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	X
			Morally Bad	+	*	
Study 2a	Normativity	Free Will Attributions	Morally Good	-	*	X
			Morally Bad	0		
		Free Will Beliefs	Morally Good	+		
			Morally Bad	+		
Study 2b	Normativity	Free Will Attributions	Morally Good	-	*	X(marg)
			Morally Bad	-	*	
		Free Will Beliefs	Morally Good	0		
			Morally Bad	0		

Study 2a	Willpower	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*(marg)	
		Free Will Beliefs	Morally Good	+	*	X(marg)
			Morally Bad	+		
Study 2b	Willpower	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	0		
		Free Will Beliefs	Morally Good	+	*	X
			Morally Bad	0		
Study 2a	Punish/Reward Response	Free Will Attributions	Morally Good	+	*	
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	
Study 2b	Punish/Reward Response	Free Will Attributions	Morally Good	+	*	X
			Morally Bad	+	*	
		Free Will Beliefs	Morally Good	+	*	
			Morally Bad	+	*	

Note. Direction indicates direction of relationship ($b \leq .01$ considered 0). Significant indicates whether the relationship is statistically significant at $p < .05$ (marg = $p < .10$). An X in the Stronger Condition column indicates that there was a significantly stronger relationship between the evaluation variable and the free will judgment in that condition than the other.

6.3 Discussion

Study 2b replicated the effects of Study 2a with an increased sample size and picked up similar patterns of results across more of the evaluation variables and free will beliefs (see Table 6 for a summary of results of both Studies 2a and 2b). Specifically, free will judgments for morally bad actions were predominantly predicted by affective retributive responses. The strongest predictors of free will judgments for morally bad actions were desires to punish, moral outrage, and perceived severity of the morally bad act. There was some indication that higher response utility ratings and lower normativity ratings predicted free will judgments for morally bad actions, but these relationships were generally weak and inconsistent across the two studies.

Across both studies, all of the six evaluation variables tended to better predict free will judgments within the morally good condition than the morally bad condition. Ratings of the generosity of the morally good act, stronger reward motives, and feelings of being morally uplifted were all stronger predictors in the morally good condition than their counterparts in the morally bad condition. Furthermore, perceived response utility, perceived normativity, and perceived willpower consistently predicted free will judgments in the morally good condition only. Thus free will attributions to morally positive actions were more sensitive to all relevant evaluations of the act including the affectively-charged evaluations (deservingness, desires to reward, emotional reaction), and the less affectively-charged evaluations of the act (utility, normativity, and willpower).

When exposed to morally bad actions, participants seemed to experience simple and strong punitive responses, which led them to perceive such actions as more freely performed. For morally good actions, a variety of features seemed to influence free will judgments, some affectively charged and some not. Of course, this specific set of findings was not predicted at the outset of this project. However, they are consistent with our general contention that responsibility judgments for morally bad actions are driven primarily by affective retributive responses, whereas free will judgments for morally good actions are more sensitive to a variety of relevant features. Furthermore, that Study 2b very closely replicated the results of Study 2a increases confidence that the present results are real and replicable.

7. General Discussion

A rather large body of work has demonstrated that people attribute more responsibility to bad actions than good and neutral ones. This created the general impression that people are only motivated to attribute responsibility to harmful actions, and not helpful ones. The present

findings have upheld that people ascribe the most responsibility to bad actions — but they also show that people ascribe responsibility to good actions. In particular, people see morally good actions as more freely performed than morally neutral ones.

Studies 2a and 2b uncovered a number of factors that help distinguish the processes by which people come to ascribe responsibility for morally negative and morally positive actions. Specifically, free will judgments for morally bad actions were predominantly influenced by affective punitive motives (i.e., desires to punish, moral outrage, and perceived severity of the act). These results are consistent with prior work demonstrating that people punish retributively with relatively little concern for the deterrent benefits of punishment. The negative emotional experience of anger may crowd out other more rational or pragmatic considerations. In contrast, free will ascriptions for morally positive actions were sensitive to all measured evaluations, including those relevant to affective reward responses (e.g., desires to reward, feelings of being morally uplifted, and perceived generosity of the act), but also less affectively-charged considerations (the likely utility of reward, normativity ratings, and perceived willpower). The positive emotional response of feeling morally uplifted may expand awareness and allow for more careful consideration of these less affectively-charged components.

Though affective punitive motives were the strongest predictors of free will ascriptions to the morally bad actor, the relationships between the emotional response items and free will judgments were actually weaker in the morally bad condition than the morally good condition. These results might suggest that even relatively low levels of affective punitive feelings can increase free will judgments for morally bad actions. These results may help explain why bad actions are generally seen as freer than good ones. People may have a relatively low threshold for ascribing free will to immoral actors. When exposed to morally bad actions, strong punitive

motives may overwhelm one's ability or willingness to consider the nuances of a particular situation. Because people are highly sensitive to harmful acts, contextual features such as whether punishment serves any deterrent purpose may not be considered. In the morally good condition, all measured evaluations (including affective responses) predicted higher free will judgments to an even stronger degree, suggesting that higher levels of these reactions may be required to perceive morally good actions as free. In other words, people might have stricter criteria for ascribing free will to morally positive actions, leading to generally weaker tendencies to see such actions as freely performed (as compared with morally bad acts).

7.1 Affect vs. Utility

Ultimately, humans are inclined to punish because punishment deters harmful behavior (e.g., Fehr & Gächter, 2002; Henrich et al., 2006). Proximately, however, people are inclined to punish to satisfy desires for retribution, and people do not demonstrate much concern for whether a particular punishment would deter bad behavior (e.g., Carlsmith et al., 2002; Crockett et al., 2014). The present results suggest that ascriptions of responsibility for morally bad actions are similarly driven by affective desires for punishment. To our knowledge, this distinction between affective and utilitarian motives for punitive behavior has not been made for reward behavior. The results of Studies 2a and 2b suggest the possibility that people may ascribe responsibility to morally good actions due to both affective desires to reward, but also more pragmatic considerations such as whether rewards will encourage future good behavior, and other contextual features of the act such as how counternormative the morally good behavior was and how much willpower was required to perform the morally good action.

How much are people's responses to moral actions influenced by the utility of reward or punishment? Studies 2a and 2b manipulated the utility explicitly. The utility manipulation was

successful, and, consistent with previous findings, did not influence punishment decisions. In contrast, utility did produce some significant effects on reward decisions, albeit inconsistently. In Study 2a, higher utility of reward increased desires to reward only in the high deservingness condition. In Study 2b, higher utility of reward increased desires to reward in both the high and low deservingness conditions. Furthermore, the utility manipulation had no main or interactive effect on free will judgments. The regression analyses, which included participants' *perceptions* of the utility of punishment and reward, provided further suggestive evidence that free will judgments for morally positive actions are sensitive to utilitarian concerns (i.e., perceived utility of reward predicted higher free will attributions and beliefs in both Studies 2a and 2b), but perceived utility of punishment also predicted higher free will judgments for morally negative actions to some extent (though this effect was generally weaker and did not emerge for free will attributions in Study 2b). Future work might employ different manipulations of utility to try to replicate the present results experimentally.

7.2 Why Rewards May be More Sensitive to Context

Selfish behavior within a society can lead to collective disaster for the entire social group (e.g., Kerr, 1983), thus reducing harmful actions is the top priority for benefiting the self and society. It is not surprising then, that exposure to morally bad actions elicits strong desires to punish and commands the strongest reactions. Rewarding good actions is less urgent, but also has value (Fehr et al., 1997). Thus, as the present results demonstrate, people are sensitive to morally good actions as well. The greater context-sensitivity of responses to morally good than morally bad actions may reflect the greater need to discriminate. Punishing a harmful action may be useful regardless of whether the action was counternormative: People are inclined to behave in normative ways (e.g., Goldstein, Cialdini, & Griskevicius, 2008; Nolan, Schultz, Cialdini,

Goldstein, & Griskevicius, 2008), and so punishment may be needed to deter harmful normative behavior. In contrast, reward may not be needed to encourage helpful normative behavior. For example, a person might regularly speed on the highway because most others do so, but not if he or she sees a police car and fears receiving a ticket. In contrast, a person might hold the door open for a stranger when entering a building because most others do so, and would do so regardless of whether he or she received a reward for doing so. Rewards would not increase normative good actions, because people are inclined to perform them even without rewards (e.g., Goldstein et al., 2008; Nolan et al., 2008). Therefore, normativity may matter more when considering rewards, whereas punishment may be useful regardless of whether an action is normative.

Similarly, punishing a harmful action may be useful regardless of whether the action was performed willfully. Punishing even unintended and accidental harmful actions, such as negligent behavior, may increase vigilance and reduce the risk of careless behavior in the future. If a person knows that they might be punished for accidentally bringing about a harmful outcome, they might seek to avoid behaviors that could potentially cause harmful outcomes. In contrast, it may not be particularly useful to reward unwilled positive actions, such as happy accidents and positive side-effects. People tend to perform their desired actions regardless of whether they anticipate unintended positive side-effects, and so rewards might do little to encourage such behaviors.

Punishment may be helpful for deterring bad behavior and bad outcomes regardless of an agent's inclinations, desires, or intentions. In contrast, reward may be helpful for encouraging good behavior and good outcomes only when the actor would not have pursued the good behavior otherwise. Thus, careful evaluation of context and motives for morally significant

actions may be more useful when allocating rewards. Furthermore, because people are generally cooperative on a day-to-day basis, it may be impractical to reward all good actions. It might therefore behoove individuals to evaluate good actions and the implications of reward more carefully. Future work might seek to manipulate normativity and willpower to test the significance of these variables experimentally.

7.3 The Function of Free Will Beliefs

The present work also speaks to the function of free will beliefs. A large body of work has demonstrated that belief in free will bolsters feelings of moral responsibility (e.g., Baumeister et al., 2009; Protzko et al., 2016; Shariff et al., 2014; Vohs & Schooler, 2008). Furthermore, people bolster their own belief in free will so as to justify holding others morally responsible (Clark et al., 2014; 2017). Another analysis suggests that there may be individual differences such that people who are particularly inclined to see a variety of negative actions and outcomes (e.g., homelessness, drug addiction) as having moral significance are also more inclined to believe that people have freedom and control over those actions and outcomes (Everett et al., 2017). Relatedly, people tend to downplay their own free will when considering the morally negative consequences of their own behavior (Vonasch, Clark, Lau, Vohs, & Baumeister, 2017). Such results seem to suggest that free will is largely about moral responsibility.

However, some recent work suggests that people ascribe free will to individuals to increase accountability more generally so as to signal that the person is capable of altering their behavior in future situations (Alquist et al., 2015; Feldman et al., 2016; Feldman & Chandrashekar, 2017). In other words, people might ascribe free will not only to increase *moral*

responsibility, but also to increase accountability across both moral and amoral domains to maximize positive outcomes in the future.

Given the strength and pervasiveness of free will beliefs, both general free will beliefs and individual ascriptions of free will likely have many antecedents and many functions. One of the strongest driving factors is likely the powerful subjective experience of conscious will (e.g., Wegner, 2002, 2003). These feelings of personal free will, and the belief that other humans have free will, both are important if not indispensable parts of human existence and social group functioning. Further work is needed to fully understand how these beliefs influence individual behavior, and how these beliefs arise through both motivated cognition and experiential observation.

7.4 Is Good Stronger than Neutral?

Whereas previous work has found that "bad is freer than good," the present results demonstrate that good is freer than neutral. However, it remains unknown whether this pattern of findings would extend beyond responsibility attributions to the broader phenomenon of "bad is stronger than good." Is bad stronger than good, but good stronger than neutral? Perhaps attesting to the phenomenon that bad is stronger than good, large bodies of work have compared bad actions and outcomes to 1) good actions and outcomes, 2) neutral actions and outcomes, and 3) less bad actions and outcomes, but little work compares good and neutral. Future work testing positive-negative asymmetries should consider including neutral or control conditions to more fully understand the significance of valence in how people understand and experience the world.

8. Conclusion

Societies dedicate far more resources toward punishing rule breakers than toward rewarding rule followers. It is also evident that individuals care a great deal more about holding

others accountable for their harmful behaviors than their helpful ones. This rather obvious moral asymmetry has led researchers to focus almost exclusively on how individuals judge morally bad actions at the expense of understanding responsibility judgments for morally good actions. However, the quality of social life can be improved both by reinforcing positive actions and discouraging harmful ones. On that basis, one might reasonably expect that people would care about accountability for positive actions too. The present research confirms that people do care about responsibility for good deeds: People ascribe more free will to morally positive actions than neutral ones. Indeed, our participants not only attributed free will to virtuous actions — they did so in a cognitively sophisticated manner that was sensitive to various specific features of the action (the normativity and required willpower) and of the reward process (whether rewards would encourage good behavior). Though people perceive harmful actions as more freely performed than helpful ones, they still perceive helpful actions as more freely performed than neutral ones. Our findings augment previous work and correct the false impression that people only care about assigning responsibility and free choice to bad actions. The condemnation of morally bad actions is simple and strong, whereas the recognition of morally good actions is weaker and more sensitive to context.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556-574.

Alquist, J. L., Ainsworth, S. E., Baumeister, R. F., Daly, M., & Stillman, T. F. (2015). The making of might-have-beens: Effects of free will belief on counterfactual thinking. *Personality and Social Psychology Bulletin*, *41*, 268-283.

Angrist, J., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review*, *99*, 1384-1414.

Ariely, D., Bracha, A., & Meier, S. (2007). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *The American Economic Review*, *120*, 544-555.

Bargh, J. A. (2008). Free will is un-natural. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 128–154). Oxford, England: Oxford University Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323-370.

Baumeister, R. F., Clark, C. J., & Luguri, J. (2015). Free will: Belief and reality. In A. R. Mele, (Ed.). *Surrounding free will: philosophy, psychology, neuroscience* (pp. 49-71). Oxford, England: Oxford University Press.

Baumeister, R. F., Masicampo, E. J., & DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, *35*, 260-268.

Baumeister, R. F., & Newman, L. S. (1994). Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin, 20*, 3-19.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333-342.

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst*, *24*, 1-44.

Card, N. A. (2012). *Applied meta-analysis for social science research.* New York, NY: Guilford

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*, 437-451.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*, 119-137.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83,* 284-299.

Clark, C. J., Bauman, C. W., Kamble, S. V., & Knowles, E. D. (2016). Intentional sin and accidental virtue? Cultural differences in moral systems influence perceived intentionality. *Social Psychological and Personality Science*, *8*, 74-82.

Clark, C. J., Baumeister, R. F. & Ditto, P. H. (2017). Making punishment palatable: Belief in free will alleviates punitive distress*. Consciousness and Cognition, 51,* 193-211.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, *106*, 501-513.

Clark, C. J., Chen, E. E. & Ditto, P. H. (2015). Moral coherence processes: Constructing culpability and consequences. *Current Opinion in Psychology, 6,* 123-128.

Clutton-Brock, T. H. & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373,* 209-216.

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*, 2279-2286.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*, 448-474.

Cushman, F. (2013). The role of learning in punishment, prosociality, and human uniqueness. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution (Life and mind: Philosophical issues in biology and psychology)* (pp. 333–372). Cambridge, MA: MIT Press.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108*, 281-289.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125,* 627-668.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568-584.

Ditto, P. H., Pizarro, D. A., Tannenbaum, D. (2009). Motivated moral reasoning. In B. H. Ross (Series Ed.) & D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Psychology of learning and motivation, Vol. 50: Moral judgment and decision making* (pp. 307-338). San Diego, CA: Academic Press.

Everett, J. A. C., Clark, C. J., Luguri, J. B., Earp, B. D., Ditto, P. H., & Shariff, A. F. (2017). Political differences in free will belief are driven by differences in moralization. *PsyArXiv.* Available at h5ps://osf.io/preprints/psyarxiv/nx9rj.

Fabes, R. A., Fultz, J., Eisenberg, N., May-Plumlee, T., & Christopher, F. S. (1989). Effects of rewards on children's prosocial motivation: A socialization study. *Developmental Psychology*, *25*, 509-515.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

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.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63-87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140.

Fehr, E., & Gächter, S. (2005). Human behaviour: Egalitarian motive and altruistic punishment (reply). *Nature*, *433*, E1-E2.

Fehr, E., Gächter, S. & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica, 65,* 833-860.

Feldman, G., & Chandrashekar, S. P. (in press). Laypersons' beliefs and intuitions about free will and determinism: New insights linking the social psychology and experimental philosophy paradigms. *Social Psychological and Personality Science.*

Feldman, G., Wong, K. F. E., & Baumeister, R. F. (2016). Bad is freer than good: Positive–negative asymmetry in attributions of free will. *Consciousness and Cognition*, *42*, 26-40.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *American Psychologist, 56*, 218–226.

Fredrickson, B. L. (2013). Positive emotions broaden and build. *Advances in Experimental Social Psychology*, *47*, 53.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion*, *19*, 313-332.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*, 472-482.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, *25*, 191-209.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535-549.

Haidt J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J. (2003). Elevation and the positive psychology of morality. In C. L. M. Keyes & J. Haidt (Eds.), *Flourishing: Positive psychology and the life well-lived* (pp. 275–289). Washington, DC: American Psychological Association.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York, NY: The Guilford Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795-815.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., ……., Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science, 327,* 1480-1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767-1770.

Irwin, K. (2009). Prosocial behavior across cultures: The effects of institutional versus generalized trust. *Advances in Group Processes*, *26*, 165-198.

Isen, A. M., Rosenzweig, A. S., & Young, M. J. (1991). The Influence of Positive Affect on Clinical Problem-Solving. *Medical Decision Making, 11*, 221–227.

Kanouse, D. E., & Hanson. L. R. (1971). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.) *Attribution: Perceiving the causes of behavior* (pp. 47-62). Morristown, NJ: General Learning Press.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681-706.

Kerr N. L. (1983). Motivation losses in task-performing groups: A social dilemma analysis. *Journal of Personality and Social Psychology, 45,* 819-828.

Kerr, N.L., & Bruun, S. (1983). The dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology, 44,* 78–94.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309-324.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology*, *2*, 441-448.

Kubanek, J., Snyder, L. H., & Abrams, R. A. (2015). Reward and punishment act as distinct factors in guiding behavior. *Cognition*, *139*, 154-167.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480-498.

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, *76*, 225-237.

Latané B., Williams K., Harkins S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822-832.

Lewicka, M., Czapinski, J., & Peeters, G. (1992). Positive‐negative asymmetry or 'When the heart needs a reason'. *European Journal of Social Psychology*, *22*, 425-434.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*, 529-566.

Liu, B. S., & Ditto, P. H. (2013). What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science*, *4*, 316-323.

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, *9*, 343-351.

Monroe, A. E., Ainsworth, S. E., Vohs, K. D., & Baumeister, R. F. (2017). Fearing the future? Future-oriented thought produces aversion to risky investments, trust, and immorality. *Social Cognition*, *35*, 66-78.

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*, 636-653.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner 1, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*, 561-584.

Nelissen, R. M., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*, 543-553.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind and Language*, *19*, 473-502.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*, 663-685.

Nietzsche, F. (1954). *Twilight of the idols* (W. Kaufmann, Trans.). New York, NY: Penguin Books. (Original work published 1889).

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, *34*, 913-923.

Orbell, J., & Dawes, R. (1981). Social dilemmas. In G. Stephenson, & J. H. Davis (Eds.), *Progress in applied social psychology* (pp.37-65.). London, England: Wiley.

Paulhus, D. L., & Carey, J. M. (2011). The FAD–Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, *93*, 96-104.

Pohling, R., & Diessner, R. (2016). Moral elevation and moral beauty: A review of the empirical literature. *Review of General Psychology*, *20*, 412-425.

Protzko, J., Ouimette, B., & Schooler, J. (2016). Believing there is no free will corrupts intuitive cooperation. *Cognition*, *151*, 6-9.

Rasmussen, E. B., & Newland, M. C. (2008). Asymmetry of reinforcement and punishment in human choice. *Journal of the Experimental Analysis of Behavior*, *89*, 157-167.

Reeder, R. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology, 44,* 736-745.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296-320.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., & Sirker, S. (2010). Is belief in free will a cultural universal?. *Mind & Language*, *25*, 346-358.

Scurich, N., & Shniderman, A. (2014). The Selective Allure of Neuroscientific Explanations. *PLoS ONE*, *9*, 1-6.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., ... & Vohs, K. D. (2014). Free Will and Punishment A Mechanistic View of Human Nature Reduces Retribution. *Psychological Science*, *25*, 1563-1570.

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology*, *22*, 435-452.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543-545.

Szolnoki, A., & Perc, M. (2010). Reward and cooperation in the spatial public goods game. *EPL*, *92*, 38003.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). Tutorial: How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35,* 215-247.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will encouraging a belief in determinism increases cheating. *Psychological Science*, *19*, 49-54.

Vonasch, A. J., Clark, C. J., Lau, S., Vohs, K. D., & Baumeister, R. F. (2017). Ordinary people associate addiction with loss of free will. *Addictive Behaviors Reports*, *5*, 56-66.

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, *3*, 73-79.

Warneken, F., & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology, 44,* 1785-1788.

Wegner, D.M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences, 7*, 65-69.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*, 480-492.