

Not so Motivated After All? Three Replication Attempts and a Theoretical Challenge to a Morally-Motivated Belief in Free Will

Andrew E. Monroe^{1*} and Dominic Ysidron²
¹Appalachian State University, ²Ohio University

Abstract

Free will is often appraised as a necessary input to for holding others morally or legally responsible for misdeeds. Recently, however, Clark and colleagues (2014), argued for the opposite causal relationship. They assert that moral judgments and the desire to punish motivate people's belief in free will. In three experiments—two exact replications (Studies 1 & 2b) and one close replication (Study 2a) we seek to replicate these findings. Additionally, in a novel experiment (Study 3) we test a theoretical challenge derived from attribution theory, which suggests that immoral behaviors do not uniquely influence free will judgments. Instead, our non-violation model argues that norm deviations, of any kind—good, bad, or strange—cause people to attribute more free will to agents, and attributions of free will are explained via desire inferences. Across replication experiments we found no evidence for the original claim that witnessing immoral behavior causes people to increase their belief in free will, though we did replicate the finding that people attribute more free will to agents who behave immorally compared to a neutral control (Studies 2a & 3). Finally, our novel experiment demonstrated broad support for our norm-violation account, suggesting that people's willingness to attribute free will to others is malleable, but not because people are motivated to blame. Instead, this experiment shows that attributions of free will are best explained by people's expectations for norm adherence, and when these expectations are violated people infer that an agent expressed their free will to do so.

Keywords: Free Will, Blame, Moral Judgment, Motivational Bias, Replication

*Correspondence to:
Andrew E. Monroe
Department of Psychology
Appalachian State University
222 Joyce Lawrence Ln.
Boone, NC 28608, USA
E-mail: monroea1@appstate.edu

Not so Motivated After All? Three Replication Attempts and a Theoretical Challenge to a Morally-Motivated Belief in Free Will

Over the last 20 years research on belief in free will has exploded as psychologists joined philosophers to examine how people's belief in free will affects moral behavior and moral judgment. Debates about free will's impact on moral behavior and moral judgment have repeatedly featured in the popular press (e.g., Nahmias, 2011; Overbye, 2007; Stafford, 2013), and psychologists, in particular, led the charge in studying the effects of people's (dis)belief in free will. Disbelief in free will has been associated with a host of morally-relevant outcomes, including less creativity and more conformity (Alquist, Ainsworth, & Baumeister, 2013), less gratitude (Crescioni, Baumeister, Ainsworth, Ent, & Lambert, 2015; MacKenzie, Vohs, & Baumeister, 2014), less self-control (Rigoni, Kühn, Gaudino, Sartori, & Brass, 2012) more aggression and less helpfulness (Baumeister, Masicampo, & DeWall, 2009), less counterfactual thinking (Alquist, Ainsworth, Baumeister, Daly, & Stillman, 2014), increases in situational causal attributions (Genschow, Rigoni, & Brass, 2017), and volition (Rigoni, Kühn, Sartori, & Brass, 2011).

Recent research, however, seeks to reverse the relationship between free will and morality, proposing that moral judgments inform people's belief in free will (Clark et al., 2014). If true, this would represent a coup for scholars' and laypersons' conception of free will free will and morality. In the present work, however, we argue that there are methodological and theoretical reasons to doubt a morally-motivated account of free will belief.

Motivated Free Will Belief and Its Challenges

In a recent paper, Clark et al. (2014) present a series of studies arguing that being exposed to immoral behavior (compared to a neutral behavior) causes people to inflate their belief in free will as a post-hoc justification for blaming. In their words, "We propose that the pervasive belief in free will partially flows from a desire for moral responsibility in order to justify punishing others for their antisocial behaviors." (Clark et al., 2014, p. 503). Moreover, Clark goes beyond other motivated cognition accounts, which envision negative behaviors biasing judgments of acts or agents (e.g., Alicke, 1992; Ditto, Pizarro, & Tannenbaum, 2009; Knobe, 2003; Tetlock et al., 2007), and instead argue that their data demonstrate that "people respond to immoral actions not merely by altering their one-time judgments about specific actions, but by shifting their broad beliefs about all humankind." (Clark et al., 2014, p. 502)

Accepting this provocative claim would invite significant and disruptive upheaval upon modern legal theory, which presumes that it is permissible and just to punish offenders' misdeeds because they had the free will to reasonably "have done otherwise" (Greene & Cohen, 2004). Similarly, this account would threaten philosophical arguments wherein free will is assumed as a necessary condition for holding people morally responsible (Aristotle, 1985; Kant, 1953; Nichols & Knobe, 2007; Sarkissian et al., 2010). The claim of a morally-motivated belief in free will, however, may rest on tenuous footing.

Replicating Clark et al (2014): A Methodological Challenge

At a methodological level, Clark's claim relies on a commonly used measure of free will belief: the Free Will Subscale of the Free Will and Determinism Scale (FAD+, Paulhus & Carey, 2011). However, three of the seven items of the Free Will Subscale ("People must take full

responsibility for any bad choices they make.” “Criminals are totally responsible for the bad things they do.” “People are always at fault for their bad behavior.”) appear to confound belief in free will with moral evaluation. Assertions that agents “must take full responsibility” or “are always at fault” are punitive judgments, not affirmations of humans’ metaphysical freedom. Clark et al. dismiss this concern, suggesting that “the fact that researchers include aspects of responsibility and blame into their operational definition of free will can be taken as additional evidence of the deep psychological association between these constructs...” (Clark et al., 2014, p. 511). However, the crux of Clark’s findings is that observing immoral behaviors increases people’s belief in free will. If the free will measure is partially measuring moral judgment, then it becomes unclear if the findings reflect people believing more in free will or merely increasing their condemnation of immoral behavior.

Clark et al. claim that analogous findings with *agent-specific free will attributions* insulate their conclusions from such criticism. However, this measure assesses only free will attributions for a specific actor (e.g., did X exercise her free will), and Clark explicitly frames the findings in more sweeping terms: “Participants are not merely attributing greater freedom, responsibility, intention, or control to the perpetrator at the specific time of the incident, but to all people in general—including the self—at all points in their lives.” (Clark et al., 2014, p. 510). Agent-specific judgments are insufficient to defend such an expansive claim; broad support from the Free Will Subscale is necessary. Yet, a *prima facie* reading of this scale suggests that nearly half of the items are imbued with moral valuation, and it is therefore unclear whether participants’ responses reflect free will beliefs, moral judgments, or some amalgam of the two. Our methodological challenge takes up this issue and tests whether the original findings replicate after removing the potentially morally contaminated items from the Free Will Subscale.

Theoretical Challenge

In addition to the methodological challenge, there is a theoretical reason to doubt the claim that immoral behaviors uniquely motivate free will beliefs. Clark’s claim is that people’s desire to punish immoral behavior motivates their belief in free will. However, seminal social psychological work on attribution (Heider, 1944; Jones, 1979, 1990; Jones & Davis, 1965) suggests an alternative, and perhaps broader mechanism at work. Specifically, these theorists argue that deviations from established norms lead to greater internal (i.e., personal) attributions. For example Heider’s (1944) “contrast effects” suggest that if a person’s present behavior contrasts with an established past pattern or a strong situation, this violation of expectation produces strong inferences about the individual (e.g., motives, dispositions). Similarly, Jones and Davis’s (1965) theory of correspondent inferences argues that behavior is perceived as most diagnostic when it is unexpected or when it deviates from a norm (Jones, 1979; Jones & Davis, 1965). When a behavior violates people’s schematic expectations—for example, a Northern college student arguing in favor of segregation (Jones & Harris, 1967)—or when a behavior violates experimenter-manipulated expectations (Jones, Worchel, Goethals, & Grumet, 1971), perceivers draw strong inferences about an actor’s goals, attitudes, or character.

These findings imply a broader explanation for how behavior might influence free will ascriptions. Behaving immorally (e.g., cheating, stealing, intentionally harming) violates widely accepted norms of social behavior, and one would therefore expect people to attribute more desire, commitment, or choice — in a word, more free will (see Monroe & Malle, 2010) — to agents who behave immorally relative to a neutral scenario. However, past attribution research demonstrates that although immoral behavior may be *sufficient* to trigger increased ascriptions of

volition, immorality is not *necessary* for these judgments. Instead, drawing on classical work on contrast effects and correspondent inferences, we argue that norm violations of *any* kind—good, bad, or strange—should increase personal attributions (i.e., whether a person made a choice, could have done otherwise, or had free will). This account explains Clark’s effects, but (1) it does so without positing the additional step of a motivated desire to punish, and (2) it makes a more expansive prediction that free will ascriptions should be sensitive to *any* type of substantive norm violation.

Experiments and Predictions

We present four preregistered experiments testing the effects of immoral (Studies 1-2b) and norm-violating behavior (Study 3) on a general belief in free will (Studies 1-3) and agent-specific ascriptions of free will (Studies 2a-3). Studies 1 and 2b are exact replications of Clark et al.’s (2014) Studies 1 and 2, respectively, and our Study 2a is a close replication of Clark et al. (2014) Study 2 (we use identical measures and procedures, but a different sample population). In these experiments we seek to replicate the finding that people inflate their belief in free will after being exposed to another agent’s immoral behavior and that the desire to punish immoral actors mediates the relationship between behavior and people’s belief in free will.

Study 3 is a novel experiment that tests our theoretical challenge by contrasting Clark et al.’s (2014) motivated model with a norm-violation model. This experiment tests whether increases in free will beliefs are a unique response to immoral behaviors, or if these effects are just as strong for other norm deviations. We predict that an agent who deviates from expectations by behaving in a morally negative, positive, or simply strange manner will be viewed as having more free will than a neutral control. Moreover, in line with foundational social psychological research (Heider, 1944; Jones, 1979; Jones & Davis, 1965) and past work on the folk concept of free will (Monroe, Brady, & Malle, 2017; Monroe & Malle, 2010, 2014; Stillman, Baumeister, & Mele, 2011), we predict that inferences of desire will mediate the relationship between behavior and free will judgments.

All of our materials, syntax, and data publicly available via the OSF (osf.io/rwe2t). We preregistered our sample size, hypotheses, and data analysis plans, and we report all of our measures.

Study 1

Methods

Participants

We conducted an a priori power analysis using the effect size from Clark et al. (2014) Study 1 ($d = 0.43$) as our baseline. We assumed an effect size of $d = 0.4$ and computed the required sample size to achieve 95% power (G*Power, independent samples t-test). The analysis showed a required sample of 328 participants; however, we elected to oversample to 400 participants in case of attrition.

In total, we recruited 406 participants ($M_{\text{age}} = 35.78$ years, $SD = 12.12$) from Amazon Mechanical Turk (AMT), and paid them \$0.25 each. Of the total sample, 231 were female. The majority identified as White/Caucasian ($n = 290$), with smaller numbers identifying as Asian/Asian American ($n = 35$), African American ($n = 29$), Latin/Hispanic ($n = 25$), or multiethnic ($n = 21$). Participants were moderately religious ($M = 2.62$, $SD = 1.54$; 1 = not at all

religious; 5 = very religious) and politically moderate ($M = 3.72$, $SD = 1.78$; 1 = very liberal; 7 = very conservative).

Design and Procedure

Participants completed the study online. They were told that they were participating in a study about memory and were randomly assigned to one of two conditions: immoral behavior ($N = 205$) or morally-neutral behavior ($N = 201$). In the immoral behavior condition participants read a news story entitled “*Nation rocked by ‘jailing kids for cash’ scandal*,” which described a Pennsylvania judge who accepted bribes to sentence minors to a juvenile detention center in order to increase their profits. In the morally neutral condition participants read a news story entitled “*Luzerne County school district starts superintendent search*” which described a Pennsylvania school district hiring a new superintendent (See Supplementary materials).

After reading one of these news stories, participants were asked to complete a series of “personality scales.” Identical to Clark, et al. (2014), participants first completed a short form of the Social Desirability Scale (Reynolds, 1982) in order to avoid raising suspicion regarding the goals of the study. The SDS contained four statements (e.g., It is sometimes hard for me to go on with my work if I am not encouraged.) rated on a 5-point Likert scale (1 strongly disagree, 5 strongly agree). Afterwards, participants reported their free will beliefs using the free will subscale of the FAD+ (Paulhus & Carey, 2011). This subscale contains seven items designed to measure people’s belief in free will. Each item was measured on a 5-point Likert scale (1 strongly disagree, 5 strongly agree). The moralized version of the free will subscale contained all seven items of the scale; whereas the non-moralized version omitted items 2, 4, and 7 (“People must take full responsibility for any bad choices they make.” “Criminals are totally responsible for the bad things they do.” “People are always at fault for their bad behavior.”). After these measures, participants completed a short demographic form and were debriefed.

Results

Two preregistered hypotheses guided this study: (1) We predicted that the moral valence of a target’s behavior (immoral vs. neutral) would increase *moralized* free will beliefs (measured with the free will subscale of the FAD+). (2) We predicted no significant effect of the moral valence of a target’s behavior (immoral vs. neutral) on non-moralized free will beliefs (measured with the free will subscale of the FAD+ omitting the three moralized items).

Both the moralized and the non-moralized version of the free will subscale demonstrated sufficient reliability ($\alpha = .81$ and $\alpha = .71$, respectively). However, the morality manipulation did not significantly affect free will beliefs on either measure. Moralized free will beliefs showed virtually identical patterns in the immoral ($M = 3.72$, $SD = 0.66$) and neutral behavior conditions ($M = 3.71$, $SD = 0.68$), $t(404) = -0.095$, $p = .925$, $d = 0.009$, 95% CI [-0.204, 0.185]. Similarly, non-moralized free will belief showed no differences between the immoral ($M = 3.64$, $SD = 0.72$) and the neutral behavior conditions ($M = 3.68$, $SD = 0.74$), $t(404) = 0.434$, $p = .664$, $d = 0.043$, 95% CI [-0.151, 0.238].

Discussion

In an exact replication of Clark, et al.’s (2014) Study 1, we found no evidence for the claim that people increase their belief in free will following exposure to another person’s immoral behavior. It is important to note that our failure to replicate the original finding was not

due to problems with the construction of the FAD+ as we initially hypothesized. In contrast, exorcising what we believed to be morally tainted items from the free will subscale did not affect any of the results. Instead, there was simply no discernable effect of immoral behavior on free will beliefs.

Whereas assertions based on a null findings are difficult, we specifically designed our study so that we could make inferences from the data regardless of the results. Our a priori power analysis was based on the effect size from the original study; we set a stringent criterion for achieved power (95%), and we oversampled by 10%. Indeed, a sensitivity power analysis indicated that we had sufficient power to detect effects as small as $d = .35$. Additionally, we conducted an exact replication of the original experiment, using the same sample group, stimuli, and measures in Clark, et al. (2014). Thus, these data suggest no evidence that immoral behavior motivates people's general belief in free will. However, as any result from a single study should be regarded tentatively, we performed two additional replication experiments. Further, Studies 2a and 2b take up Clark et al.'s (2014) prediction that a desire to punish mediates the relationship between behavior and free will belief.

Study 2

Two central premises underlie the claim of morally motivated free will beliefs. First is the assertion that exposing perceivers to immoral actions engenders stronger beliefs in free will. Second, immoral behaviors should prompt a desire to punish, and this desire mediates the relationship between the moral valence of behavior and free will belief. Whereas Study 1 casts doubt on the first premise, Studies 2a and 2b provide further tests of this prediction. Additionally, Studies 2a and 2b test the second premise by providing a close (Study 2a) and an exact (Study 2b) replication of Clark et al. (2014) Study 2, wherein they test for mediation. The methods and materials for Studies 2a and 2b were identical, except that Study 2a used an internet sample from Amazon Mechanical Turk (AMT), and Study 2b used a University student sample (identical to Clark et al., 2014).

We focus on three preregistered hypotheses in Studies 2a and 2b.¹ We predict that (1) the moral valence of a target's behavior (immoral vs. neutral) will increase free will beliefs; (2) the moral valence of a target's behavior will increase target-specific free will attributions, and (3) that punishment judgments will mediate the relationship between the moral valence manipulation and free will beliefs and target-specific free will attributions.

Methods

We conducted an a priori power analysis (G*Power, independent samples t-test) based on Clark et al. (2014) Study 2. The original study reports two effect sizes corresponding to general beliefs in free will ($d = 0.47$) and agent-specific free will attributions ($d = 0.51$). We assumed a slightly more conservative effect size estimate ($d = 0.4$) and computed the required sample size

¹ As in Study 1 we tested the effect of the morality manipulation on the full free will subscale and a reduced non-moralized version of the scale. However, as in Study 1 the pattern of results for the full scale and the non-moralized version were identical. We therefore focus on the results from the full free will subscale; results for the non-moralized scale are reported in the supplementary materials.

to achieve 95% power. The analysis indicated a required sample of 328 participants per study. We oversampled by setting our stopping rule at 400 participants for Studies 2a and 2b. Separate samples were recruited from AMT (Study 2a: $N = 401$) and from a Psychology undergraduate subject pool at a mid-size University (Study 2b: $N = 397$).

Study 2a Participants

We recruited 401 participants ($M_{\text{age}} = 36.02$ years, $SD = 11.49$) from AMT, and paid them \$0.25 each. Of the total sample, 229 were female. A majority identified as White/Caucasian ($n = 285$), with fewer participants identifying as Asian/Asian American ($n = 35$), African/African American ($n = 36$); Latin/Hispanic ($n = 19$), and multiethnic ($n = 21$). Participants were moderately religious ($M = 2.60$, $SD = 1.47$) and politically moderate ($M = 3.57$, $SD = 1.76$).

Study 2b Participants

We recruited 399 college students ($M_{\text{age}} = 20.57$ years, $SD = 6.64$) from a mid-size public University, and compensated them with course credit. The sample was majority female ($n = 295$) and White ($n = 344$), with smaller numbers identifying as Asian/Asian American ($n = 8$), African American ($n = 11$), Latin/Hispanic ($n = 17$) multiethnic ($n = 17$). Participants were moderately religious ($M = 3.19$, $SD = 1.35$) and politically moderate ($M = 3.84$, $SD = 1.51$).

Design and Procedure

The design and procedure of Studies 2a and 2b were identical and an exact replication of Clark et al. (2014), Study 2. The cover story, stimuli, and dependent variables were the same as those in the original study. Participants completed the study online. Participants were told they were participating in a study about memory and randomly assigned to read about either an immoral behavior (a burglary) or a morally neutral behavior (taking recycled cans).

After reading the vignette, participants responded to three items measuring agent-specific attributions of free will to the transgressor (whether the action was freely chosen, whether the actor could have made other choices, and whether the actor exercised his or her own free will) on a 7-point scale (1 not at all - 7 very much so). Afterwards, participants rated how much the transgressor should be punished for their actions (1 not at all - 7 severely). Lastly, participants completed the FAD+, a short demographic questionnaire, and were then debriefed.

Results

Study 2a

The free will subscale ($\alpha = .85$) and the agent-specific free will attribution measure ($\alpha = .80$) demonstrated sufficient reliability. Contrary to our first prediction (but mirroring our findings from Study 1), there was no effect of the morality manipulation on free will beliefs, $t(399) = -0.157$, $p = .875$, $d = 0.016$, 95% CI [-0.211, 0.180]. However, agent-specific free will attributions did show the predicted effect. Participants attributed more free will to agents who

behaved immorally compared to a neutral control behavior, $t(399) = -3.227, p = .001, d = 0.322$, 95% CI [0.125, 0.519] (See Figure 1).²

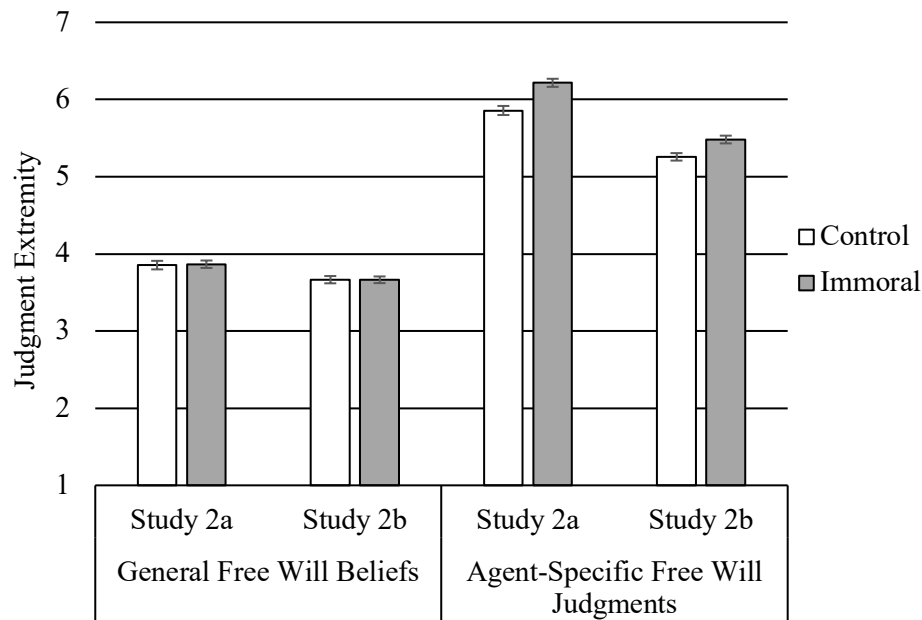


Figure 1. Two Studies find no evidence that immoral behaviors motivate general free will beliefs; however, people did attribute more free will to agents who acted immorally versus a behaving neutrally. Error bars = ± 1 SE.

Even though there was no direct effect of condition on participants' general free will beliefs, it's still possible that the mediated, indirect effect was present. We therefore conducted two mediation analyses testing whether the desire to punish mediated the relationship between condition and the (1) full free will subscale and (2) agent-specific free will attributions using bootstrapping with 10,000 samples (Hayes, 2013, model 4). Across analyses, condition significantly predicted the desire to punish, $b = 3.367, se = .160, 95\% \text{ CI } [3.053, 3.680]$.

Free will subscale. There was no direct effect of condition on moralized free will belief, $b = 0.012, se = .074, 95\% \text{ CI } [-0.132, 0.157]$, though desire to punish significantly predicted free will belief, $b = 0.088, se = .022, 95\% \text{ CI } [0.044, 0.132]$. The overall indirect effect was significant, indirect $b = 0.297, se = .077, 95\% \text{ CI } [0.151, 0.453]$; however, simultaneously entering condition and desire to punish in the model revealed a suppressor effect where the previously non-significant relationship between condition and moralized free will beliefs became significant, but negative, direct $b = -0.285, se = .104, 95\% \text{ CI } [-0.490, -0.080]$ (See Figure 2).

Agent-specific free will attributions. The predicted mediation effect emerged. Condition significantly predicted attributions of free will, $b = 0.355, se = .111, 95\% \text{ CI } [0.136, 0.574]$, and desire to punish significantly predicted free will judgments, $b = 0.080, se = .034, 95\% \text{ CI } [0.013, 0.147]$. Further, the overall indirect effect was significant, indirect $b = 0.269, se$

² Participants also recommended different amount of punishment for the agent in the control ($M = 2.62, SD = 1.94$) and the immoral condition ($M = 5.99, SD = 1.24$), $t(398) = -20.71, p < .001, d = 2.072, 95\% \text{ CI } [1.828, 2.314]$.

= .136, 95% CI [0.005, 0.537], and the direct path from condition to free will attributions was no longer significant, direct $b = 0.085$, $se = .160$, 95% CI [-0.229, 0.399] (Figure 2).

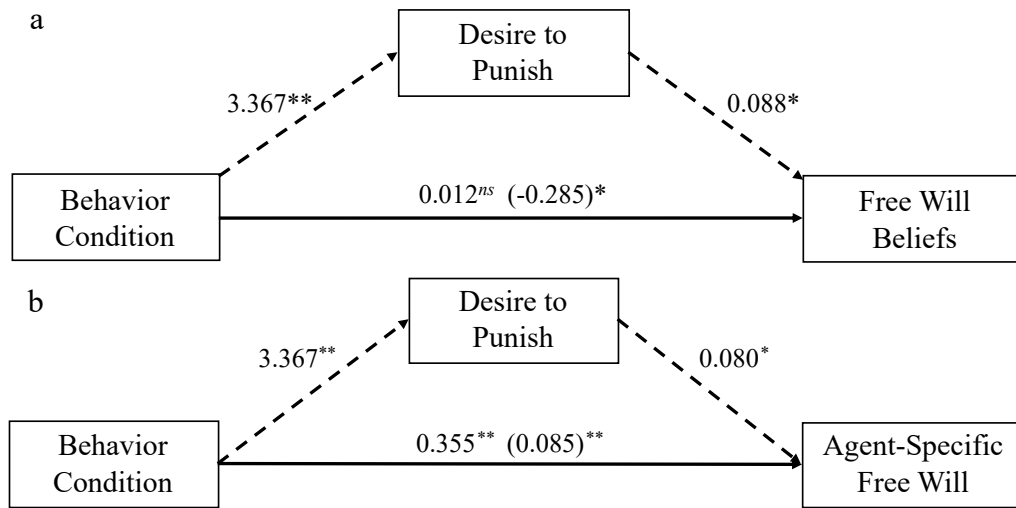


Figure 2. Mediation model demonstrating a suppressor effect. Controlling for punishment judgments, participants report less belief in free will in after reading about an immoral behavior compared to a morally neutral control behavior (panel a). Contrastingly, punishment recommendations significantly mediated the relationship between immoral behaviors and agent-specific attributions (panel b).

** $p < .001$; * $p < .05$.

Study 2b

The free will belief and agent-specific attribution measures demonstrated lower reliability compared to Study 2a ($\alpha = .77$ and $\alpha = .61$, respectively), and, as in Study 2a, there was no effect of condition on free will beliefs, $t(397) = 0.019$, $p = .985$, $d = 0.002$, 95% CI [-0.194, 0.198]. Additionally, unlike Study 2a, the effect of the moral behavior manipulation on agent-specific free will attributions failed to reach the conventional threshold for significance, $t(397) = -1.875$, $p = .062$, $d = 0.188$, 95% CI [-0.009, 0.384] (See Figure 1).³

We conducted two mediation analyses (bootstrapping with 10,000 samples; Hayes, 2013, model 4) testing whether the desire to punish mediated the relationship between condition and general free will beliefs and agent-specific free will attributions. Across analyses, condition significantly predicted the desire to punish, $b = 2.167$, $se = .184$, 95% CI [1.806, 2.529].

Free will subscale. There was no direct effect of condition on free will belief, $b = 0.002$, $se = .064$, 95% CI [-0.124, 0.129]. Desire to punish significantly predicted free will belief, $b = 0.093$, $se = .017$, 95% CI [0.060, 0.126], and the overall indirect effect was significant, indirect $b = 0.201$, $se = .042$, 95% CI [0.126, 0.291]. However, as in Study 2a, the analysis revealed that

³ As in study 2a, participants recommended clearly different amounts of punishment in the control ($M = 2.72$, $SD = 1.84$) and the immoral behavior condition ($M = 4.88$, $SD = 1.82$), $t(395) = -11.80$, $p < .001$, $d = 1.180$, 95% CI [0.966, 1.393].

entering condition and desire to punish in the model simultaneously showed a suppressor effect as the relationship between condition and free will belief became significant, direct $b = -0.199$, $se = .072$, 95% CI [-0.341, -0.057] (See Figure 3).

Agent-specific free will attributions. The mediation analysis for the agent-specific free will attributions demonstrated the predicted mediation effect. Although the initial, direct effect of condition was marginally significant, $b = 0.225$, $se = .120$, 95% CI [-0.011, 0.461], desire to punish significantly predicted free will judgments, $b = 0.188$, $se = .032$, 95% CI [0.123, 0.250], and the overall indirect effect showed significant mediation, indirect $b = 0.407$, $se = .077$, 95% CI [0.268, 0.571]. In the full model, the direct path from condition to free will attributions was no longer significant, direct $b = -0.182$, $se = .134$, 95% CI [-0.445, 0.081] (See Figure 3).

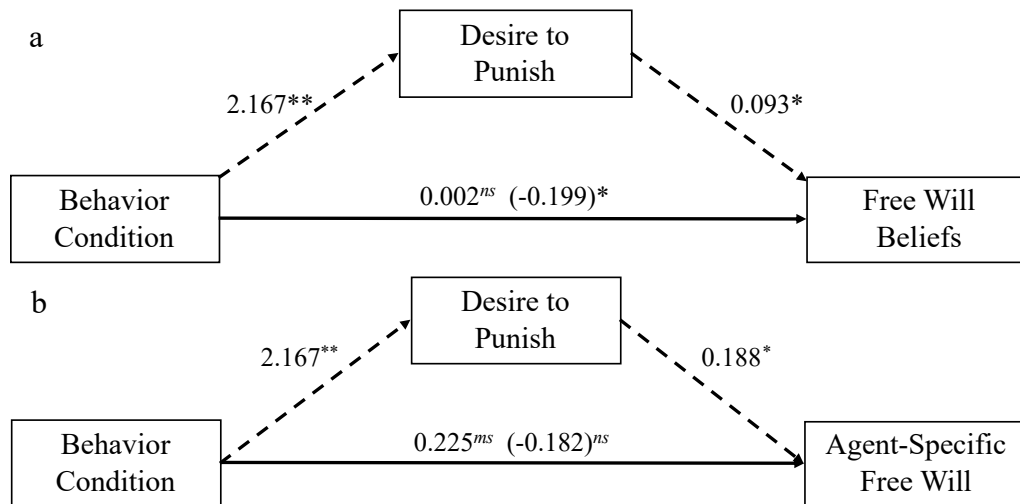


Figure 3. Replicating Study 2a, the mediation analysis revealed a significant suppressor effect. Controlling for punishment recommendations, participants believed in free will less after reading about an immoral behavior compared to a neutral control story (panel a). Replicating Study 2a, punishment recommendations significantly mediated the relationship between immoral behaviors and agent-specific free will attributions (panel b).

$^{**}p < .001$; $^*p < .05$.

Discussion

Two additional studies demonstrate scant evidence for the claim that immoral behavior motivates people to increase their general belief in free will. Neither Study 2a or 2b replicated Clark et al.'s (2014) findings regarding people's general beliefs in free will. This consistent lack of an effect suggests that the motivated belief in free will effect may be smaller than originally suggested or nonexistent. Although perhaps surprising, this null effect is consistent with recent work arguing that people's general beliefs about free will have little to do with their moral judgments (Monroe et al., 2017; Monroe & Malle, 2014). Instead, recent work suggest that people's use of the term "free will" is a shorthand for intentional agency. Once free will is unpacked into its psychological constituents—intentionality, choice, and desire—there is little left for the abstract concept of free will to predict (Monroe, Dillon, & Malle, 2014).

We did, however, find at least mixed evidence for the claim that immoral behavior increases people's willingness to ascribe free will to other agents. Study 2a showed that when people read about an agent who burgled another person's home, they believed he had more free

will than an agent who removed aluminum cans from a person's recycling bin. Study 2b, however, failed to replicate this effect, though it showed a similar descriptive pattern.

Regarding the prediction that the desire to punish mediated the relationship between observed behavior and free will belief we found support for this claim when examining the agent-specific attributions of free will. However, the mediation models predicting people's general belief in free will failed to support the original findings. In both Studies 2a and 2b, moral violations increased people's desire to punish, but they failed to affect free will beliefs. Moreover, the mediation models demonstrated a consistent suppressor effects whereby the previously nonsignificant relationship between condition and free will beliefs became significant *and negative*.⁴ That is, the analyses show that reading about moral violations increased the desire to punish, but did not directly increase free will beliefs. Moreover, when controlling for desire to punish, reading about moral violations actually *reduced* people's free will beliefs (opposite Clark et al.'s predicted pattern).

Study 3

In Study 3 we turn to our novel experiment and theoretical challenge. Clark et al. (2014) argue that people's desire to punish wrongdoing causes them to inflate their belief in free will (Clark et al., 2014; Clark, Baumeister, & Ditto, 2017). This motivated increase in free will belief is argued to justify punishment decisions (Clark et al., 2014) or to alleviate the distress people may experience from punishing others (Clark et al., 2017). Thus, one core theoretical assumption is that the effect is specific to witnessing morally negative behaviors, because only these behaviors would activate the desire to punish. Alternatively, a norm-violation account derived from attribution theory suggests that deviations from normality (bad, good, or unexpected) are likely to trigger personal attributions such as desires or dispositions (Heider, 1944; Jones & Davis, 1965; Jones & Harris, 1967). Similarly, work on people's concept of free will demonstrates that people are more likely to attribute volition to an agent when she breaks a norm compared to conforming to a norm (Monroe, Dillon, Guglielmo, & Baumeister, 2018). Thus, Clark and colleagues' (2014) motivated free will belief model and our norm violation model make different predictions regarding the types of behavior that should influence free will judgments, and the mediator of these effects.

The motivated free will belief model predicts that (1) only blameworthy behaviors should motivate free will judgments, and (2) that moral judgments (i.e., the desire to punish) will mediate the relationship between a behavior's moral valence and perceivers' free will beliefs. By contrast, the norm violation model predicts that (1) free will judgments should increase in response to any type of norm violation (i.e., praiseworthy, blameworthy, or strange behavior), and (2) that breaking a norm reveals information about agents' desires and motives (Jones & Davis, 1965; Monroe et al., 2018; Monroe, Reeder, & James, 2015; Reeder, Monroe, & Pryor, 2008) and such desire inferences should mediate the relationship between behavior and free will judgments (Table 1). To test these predictions, we created four new vignettes that describe an

⁴ One possible explanation for this suppressor effect is that there might have been a substantial collinearity problem in Clark's original manipulation (immoral vs. morally neutral) measure (punishment) measure. That is, the punishment measure may be such a close proxy for the manipulation itself, that when both are included in the mediation model, the punishment measure explains all of the variance leaving only the residual error variance for the condition manipulation to explain.

agent acting in either a blameworthy, morally neutral, morally-neutral-but-strange, or praiseworthy manner (See Supplementary Materials for stimuli norming).

Table 1.

Contrasting Predictions from Clark et al. 's (2014) Morally Motivated Free Will Model and a Norm Violation Model.

	Predicted Free Will Judgments				Predicted Mediators
	Blameworthy Behavior	Neutral Behavior	Strange Behavior	Praiseworthy Behavior	
Motivated Free Will Belief Model	↑	≈	≈	≈	Desire to Punish
Norm Violation Model	↑	≈	↑	↑	Perceived Desires

Methods

Participants

Recent large-scale replication attempts suggest that published research may overestimate true effect sizes (Open Science Collaboration, 2015), and therefore, in Study 3 we opted for a yet more conservative estimate of the motivated free will effect size ($d = 0.30$). An a priori power analysis using G*Power (Fixed effects, omnibus, one-way ANOVA; 95% power; $d = .30$, $f = .15$) revealed a total required sample size of 768 participants. We set our stopping rule at 800 participants (200 per condition).

We recruited 800 participants ($M_{\text{age}} = 36.41$ years, $SD = 11.81$) from AMT, and paid them \$0.25 each. Of the total sample, 449 (56.1%) were female. A majority identified as White/Caucasian ($n = 581$), with smaller numbers identifying as Asian/Asian American ($n = 50$), African/African American ($n = 76$), Latin/Hispanic ($n = 51$), Native American ($n = 4$), Middle Eastern ($n = 2$), multiethnic ($n = 31$). Participants were moderately religious ($M = 2.67$, $SD = 1.47$) and politically moderate ($M = 3.53$, $SD = 1.78$).

Design and Procedure

Participants were randomly assigned to read one of four vignettes, which described an agent behaving in either a blameworthy ($n = 200$), morally neutral ($n = 200$), strange, but morally neutral ($n = 200$), or praiseworthy ($n = 200$) manner (see Supplementary materials for vignettes). As in previous studies participants were told they were participating in a study about memory.

After reading the vignette, participants responded to three items measuring agent-specific free will ascriptions (identical to Studies 2a/2b), each on a 7-point scale (1 not at all - 7 very much so). Afterwards participants rated their desire to punish or reward the agent: "How much punishment or reward does [agent] deserve?" (-5 a lot of punishment; 0 neither punishment nor reward; +5 a lot of reward), and they judged the agent's desire: "How much did [agent] want to

[behavior]?” (1 – not at all, 7 very much so).⁵ Lastly, participants completed the FAD+ and then completed a short demographic questionnaire and were debriefed.

Results

Manipulation Check

We verified that people distinguished between the moral valence of the conditions as measured by people’s desire to punish or reward the agent. A univariate ANOVA demonstrated the manipulation strongly impacted people’s moral judgments, $F(3, 795), 382.7, p < .001$, partial $\eta^2 = .59$, 95% CI [0.55, 0.62]. Moral judgments were most negative in the blameworthy condition ($M = -3.22, SD = 2.54$); whereas both the control ($M = 0.49, SD = 1.30$) and the strange behavior ($M = 0.31, SD = 1.68$) were evaluated as morally neutral, and moral judgments were most positive in the praiseworthy behavior condition ($M = 2.97, SD = 1.87$). Examining differences between conditions showed that all pairwise comparisons were significant ($ps < .001$) with the exception of the comparison between the strange and the neutral conditions ($p = .34$).

Testing Two Models of Free Will Judgments

As in the previous Studies 2a and 2b we examined both people’s general belief in free will (as measured by the free will subscale of the FAD+, $\alpha = .83$) and agent-specific free will attributions ($\alpha = .85$). We conducted one-way ANOVAs for each measure, using planned contrasts to test the two theoretical models. The motivated free will belief contrast (3, -1, -1, -1) tested the prediction that free will judgments would be heightened in the blameworthy condition relative to the other three conditions. Contrastingly, the norm violation contrast (1, -3, 1, 1) predicted that all three norm violating conditions (blameworthy, praiseworthy, and strange behavior) should heighten free will judgments relative to control.

The planned contrast tests revealed that the norm violation contrast produced stronger and more consistent effects than the motivated free will contrast. Examining general free will beliefs showed that both the norm violation contrast, $t(796) = 2.46, p = .014, d = .087$, 95% CI [0.018, 0.157], and the motivated free will belief contrast, $t(796) = 2.18, p = .030, d = .077$, 95% CI [0.007, 0.147] were significant, though the norm violation model was slightly stronger.⁶ Additionally, the two models starkly diverged for agent-specific free will ascriptions. Although again both sets of contrasts were significant, the effect size for the norm violation contrast, $t(796) = 9.12, p < .001, d = .323$, 95% CI [0.252, 0.394] was twice the size of the motivated free will belief contrast, $t(796) = 5.02, p < .001, d = .178$, 95% CI [0.108, 0.248]. Indeed, as the 95% confidence intervals for these two contrasts do not overlap, the norm violation model was a statistically significantly stronger model than the motivated free will belief model (See Figure 4).

⁵ We also measured perceptions of a commonness, weirdness, and whether a behavior was perceived as breaking a norm as additional measures of the behaviors. These questions were not central to testing our hypotheses, and so we report these data in the supplementary materials.

⁶ Examining non-moralized free will beliefs showed that the norm violation contrast was significant, $t(796) = 2.54, p = .011, d = .090$, 95% CI [0.020, 0.159], whereas the motivated free will belief contrast was not, $t(796) = 1.67, p = .095, d = .059$, 95% CI [-0.010, 0.129].

Additionally, simple effects revealed mixed evidence for motivated free will belief. Although, relative to control, people attributed more free will to agents in the blameworthy condition ($p < .001$, $d = 0.82$), the blameworthy condition did not differ from praiseworthy condition ($p = .13$, $d = 0.17$), and differences from the strange condition were small ($p = .032$, $d = 0.23$). By contrast, the norm violation model was strongly supported as not only did blameworthy behaviors differ from control, but so did praiseworthy ($p < .001$, $d = 0.67$), and strange behaviors ($p < .001$, $d = 0.59$).

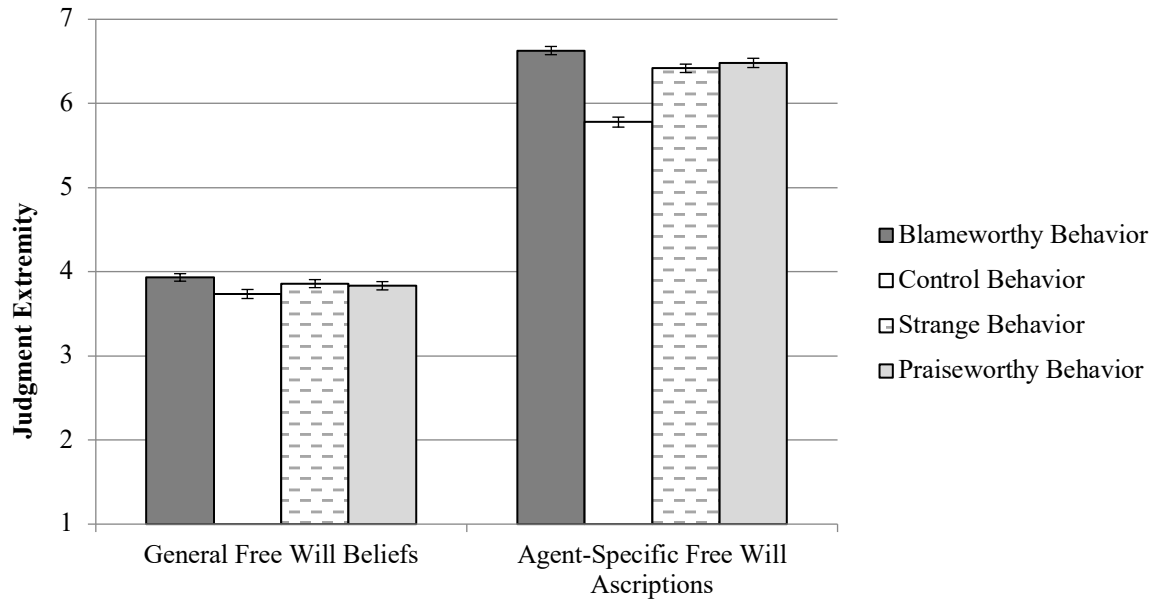


Figure 4. Blameworthy, praiseworthy, and morally-neutral-but-strange behaviors all engender stronger free will beliefs and free will judgments relative to control suggesting that norm violation, not immorality, motivates free will judgments. Error bars = ± 1 SE.

Mediation Analysis

Additionally, two models by examining whether (a) perceivers' desire to punish or (b) the agent's perceived desire to perform the behavior mediated the relationship between behavior condition and free will beliefs.⁷ Bootstrapping with 10,000 samples (Hayes, 2013, model 4) showed that the motivated free will belief mediation model (i.e., Condition \rightarrow Desire to punish \rightarrow Free will judgments) was not significant. Although condition significantly predicted desire to punish ($b = 1.91$, $se = .065$, $p < .001$) the indirect effects for all three measures of free will belief failed to demonstrate significant mediation for both general free will beliefs, indirect $b = 0.014$, $se = .022$, 95% CI [-0.028, 0.057] and for agent-specific free will attributions, indirect $b = 0.060$,

⁷ Since mediation models are affected by the coding [ordering] of the factor levels, we reordered our conditions consistent with the moral judgment manipulation check (Blameworthy, Strange, Control, Praiseworthy) so that we did not disadvantage the motivated free will model.

$se = .036$, 95% CI $[-0.009, 0.132]$.⁸ By contrast, the mediation analysis showed consistent support for the norm violation model (i.e., Condition \rightarrow Desire \rightarrow Free will judgments). Inferences of desire significantly mediated the relationship between condition and general free will beliefs, indirect $b = -0.011$, $se = .004$, 95% CI $[-0.021, -0.003]$ as well as for agent-specific free will attributions, indirect $b = -0.035$, $se = .014$, 95% CI $[-0.064, -0.011]$.

Discussion

Study 3 revealed two key findings. First, consistent with Studies 1 and 2, we demonstrate that the moral valence of agents' behaviors has little effect on people's general belief in free will. Although our expanded sample size ($n = 800$) enabled us to detect effects of moral behavior on general free will beliefs, the actual effect was substantially smaller ($p = .041$, $d = .20$) than originally reported. By contrast, we replicated previous findings that people ascribe more agent-specific free will to badly behaving agents (Clark et al., 2014). Thus, to the extent that motivated free will belief effects exist, they appear are limited to agent-specific free will ascriptions.

Secondly, Study 3 confirmed our theoretical challenge: Consistent with the norm violation account, people judged agents as having more free will whenever they broke a norm. Moreover, we found that inferences about agents' desires, rather than observers' desire to punish, mediated the relationship between observed behavior and free will judgments. These findings suggest a broader explanatory mechanism than free will being motivated by a desire to punish. The data suggest that any behavior perceived as out of the ordinary (i.e., norm violating), is viewed as being diagnostic of that agent having (or at least expressing) stronger desires and more strongly committed choices—key components of the ordinary concept of free will (Monroe et al., 2017; Monroe & Malle, 2010, 2014).

General Discussion

In their original paper, Clark et al. (2014) asserted that: “considerations of morally bad behavior would motivate people not only to attribute a greater degree of free will to the specific actor, but to believe more in the free will of people generally.” (503). Four preregistered experiments suggest that this assertion lacks empirical support. Three well-powered replications (Studies 1-2b) found no evidence that the moral valence of an actor's behavior influenced people's *general* belief in free will. Only the expanded sample in Study 3 ($n = 800$) was able to detect an effect of moral behavior on free will belief, but this effect was substantially smaller ($d = .20$) than originally reported.

The strongest evidence for Clark et al.'s (2014) claim of morally motivated free will belief derives from agent-specific free will attributions. Study 2a and Study 3 replicated the

⁸ One might argue that it is inappropriate to include the strange behavior condition in the mediation model that runs through the desire to punish because punishment is not clearly applicable to morally irrelevant behaviors. To address this concern, we reran each of the three mediation models using only the morally bad, morally neutral, and morally good conditions to predict the desire to punish and free will ascriptions. These models showed that although condition remained a significant predictor of the desire to punish ($b = 2.10$, $p < .0001$), none of the models showed evidence of significant mediation. The indirect effects predicting moralized free will beliefs ($b = -0.005$, $se = .038$, 95% CI $[-0.054, 0.062]$), non-moralized free will beliefs ($b = 0.015$, $se = .032$, 95% CI $[-0.048, 0.078]$), and agent-specific free will ascriptions ($b = 0.050$, $se = .050$, 95% CI $[-0.049, 0.145]$) were all non-significant.

original finding (and Study 2b showed a consistent, descriptive pattern) that observing an agent commit an immoral act, relative to a neutral act, causes people to attribute more free will to that agent. Thus, one conclusion that appears robust is that, relative to a neutral behavior, people attribute more free will to agents who behave immorally. However, such an effect is hardly novel in the psychological literature. Previous studies demonstrated that people are more likely to say that an agent caused (Alicke, 1992), intended (Knobe, 2003), or freely willed (Phillips & Knobe, 2009) a negative behavior more so than a neutral behavior. And further, these moral valence effects by themselves do not require a desire to punish. Recent research shows that these effects can be explained by non-moral processes such as attributional heuristics (Guglielmo & Malle, 2010), differential base rates (Uttich & Lombrozo, 2010), and structural differences in how positive and negative stimuli direct attention (Laurent, Clark, & Schweitzer, 2015).

Need for Theoretical Reinterpretation

Study 3 presents a theoretical challenge to the motivated free will belief viewpoint. Clark et al. (2014) predicate their conclusions on the claim that observing immoral behaviors activates a desire to punish the wrongdoers, and thereby causes people to inflate their belief in free will as a means to justify their desire to punish. This critical role of a desire to punish requires that the effect on free will beliefs be unique to people's response to immoral behaviors—other norm violations, such as strange or morally good behaviors, would not engender such a desire to punish. However, in three experiments (Studies 2a, 2b, 3) we found that the desire to punish failed to mediate the effect of immoral behavior on people's general belief in free will. Most critically, Study 3 revealed that norm violation more generally, not immorality specifically, explained variations in people's free will judgments. Agents who committed an immoral act, a praiseworthy act, or simply strange act were judged as having more free will than agent who performed a morally neutral act. Importantly, whereas all three norm-violating behaviors (blameworthy, praiseworthy, and strange behavior) significantly differed from the control behavior, blameworthy behaviors did not differ from the praiseworthy or the strange behavior.

Together these findings argue for a non-moral explanation for free will judgments with norm-violation as the key driver. This account explains people's tendency to attribute more free will to behaving badly agents because people generally expect others to follow moral norms, and when they don't, people believe that there must have been a strong desire to perform the behavior. In addition, a norm-violation account is able to explain why people attribute more free will to agents behaving in odd or morally positive ways. Any deviation from what is expected causes people to attribute more desire and choice (i.e., free will) to that agent. Thus our findings suggest that people's willingness to ascribe free will to others is indeed malleable, but considerations of free will are being driven by basic social cognitive representations of norms, expectations, and desire. Moreover, these data indicate that when people endorse free will for themselves or for others, they are not making claims about broad metaphysical freedom. Instead, if desires and norm-constraints are what affect ascriptions of free will, this suggests that what it means to have (or believe) in free will is to be rational (i.e., making choices informed by desires and preferences) and able to overcome constraints.

Replication Summary and Caveats

The present studies pose empirical and theoretical challenges to Clark et al.'s (2014) original thesis. Exact replications (Studies 1 & 2b), close replications (Study 2a), and novel

experiments (Study 3) find no evidence that witnessing immoral behavior causes people to increase their general belief in free will. We did replicate the finding that people ascribe more free will to agents who behave immorally (Studies 2a & 3); however, Study 3 shows that this effect is explained by a norm-violation account, not a morally motivated desire to blame.

Despite these challenges, we want to acknowledge several caveats to our findings. First, Clark et. al., (2014) measured free will in ways not captured here. Specifically, in Study 4 they demonstrate that immorality increases people's implicit belief in free will (as measured by participants' skepticism of anti-free will claims). Our data cannot speak to this implicit claim, and one might argue for preserving the motivated free will model on the basis of these implicit belief findings.⁹

A second caveat to our claims is that we do not test the societal association between crime and country-level free will belief as demonstrated in Study 5 (Clark et al., 2014). Clark and colleagues acknowledge that the study's correlational design renders it open to alternative explanations, and they therefore appeal to parsimony and their experimental findings to support their conclusion. However, if the experimental evidence fails to hold—as in our replication attempts—then it is no longer parsimonious to favor the motivated free will interpretation among the range of possible explanations.

A Recommendation for Improving Free Will Research

Research on free will has enjoyed explosive popularity over the past two decades. This expansion, however, has recently been marked by high profile failures to replicate. Several recent experiments failed to reproduce previous findings linking disbelief in free will with being more likely to cheat (Open Science Collaboration, 2015) or being less likely to behave prosocially (Crone & Levy, 2018). Further, recent studies demonstrate that commonly used free will manipulations may not be effective at shifting people's free will beliefs in the first place (Koppel, Fondacaro, & Na, in press; Monroe et al., 2017). Contrastingly, the present studies demonstrate that findings pertaining to agent-specific free will attributions are broadly replicable and are sensitive to socially-relevant factors (e.g., immorality and norm-violation). Moreover, these agent-specific judgments closely align with people's folk concept of free will, and past work demonstrates that manipulating the constituent parts of people's folk concept (e.g., choice, constraint) produce large, replicable effects on moral judgment (Monroe et al., 2017, 2014; Reeder et al., 2008; Woolfolk, Doris, & Darley, 2006).

Thus, one avenue for improving research on free will belief is to move away from focusing on people's general belief in free will to focusing on the folk concept of free will. Whereas work in experimental philosophy has been plagued by constantly contrasting findings when attempting to nail down people's metaphysical commitments regarding free will (Knobe, 2014; Nahmias, Morris, Nadelhoffer, & Turner, 2005; Nahmias, Shepard, & Reuter, 2014; Nichols, 2004; Nichols & Knobe, 2007), recent research demonstrates that people can clearly articulate the constituent parts of their free will concept as choice, intentionality, and a lack of

⁹ The original data, however, may not be robust enough to back such a claim. The reported effect on implicit free will beliefs is small ($p = .045$, $d = 0.37$, 95% CI [0.102, 0.643]). More problematically, the effect is only significant when omitting 11 participants who did not respond with a 4 or a 5 on 1- 5 scale on how seriously they filled out the survey. Including these 11 participants substantially reduced the effect size ($p = .077$, $d = 0.24$, 95% CI [-0.025, 0.500]).

coercion (Monroe & Malle, 2010, 2014; Stillman et al., 2011; Vonasch, Baumeister, & Mele, 2018), and people easily apply this concept to moral decisions to blame and punish (Monroe et al., 2017) or to determine what types of agents can be morally responsible (Monroe et al., 2014).

If researchers want to test how variations in people's belief in free will affects behavior or how perceptions of other's behavior affect free will belief, then researchers must ground such work in an accurate theory of what constitutes free will belief. Continuing to develop research on people's folk concept of free will and better integrating it to tests of behavioral outcomes may be an important foundation upon which a reliable science of free will can be built. Without this foundation, researchers may continue to produce a bevy of provocative findings, but the broader meaning of these findings will remain unclear.

References

- Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, 63, 368–378. doi:10.1037/0022-3514.63.3.368
- Alquist, J. L., Ainsworth, S. E., & Baumeister, R. F. (2013). Determined to conform: Disbelief in free will increases conformity. *Journal of Experimental Social Psychology*, 49, 80–86. doi:10.1016/j.jesp.2012.08.015
- Alquist, J. L., Ainsworth, S. E., Baumeister, R. F., Daly, M., & Stillman, T. F. (2014). The making of might-have-beens: Effects of free will belief on counterfactual thinking. *Personality and Social Psychology Bulletin*, 0146167214563673. doi:10.1177/0146167214563673
- Aristotle. (1985). *Nicomachean ethics*. (T. Irwin, Trans.). Indianapolis, IN: Hackett.
- 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. doi:10.1177/0146167208327217
- 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. doi:10.1016/j.concog.2017.03.010
- 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. doi:10.1037/a0035880
- Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., & Lambert, N. M. (2015). Subjective correlates and consequences of belief in free will. *Philosophical Psychology*, 0, 1–23. doi:10.1080/09515089.2014.996285
- Crone, D., & Levy, N. L. (2018). Are free will believers nicer people? (Four studies suggest not). *Social Psychological and Personality Science*. doi:10.17605/OSF.IO/ZPJ5X
- Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making*, The psychology of learning and motivation; Vol 50; 0079-7421 (Print); (Vol. 50, pp. 307–338). San Diego, CA US: Elsevier Academic Press.
- Genschow, O., Rigoni, D., & Brass, M. (2017). Belief in free will affects causal attributions when judging others' behavior. *Proceedings of the National Academy of Sciences*, 114, 10071–10076. doi:10.1073/pnas.1701916114
- Greene, J. D., & Cohen, J. D. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359, 1775–1785. doi:10.1098/rstb.2004.1546
- Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 36, 1635–1647. doi:10.1177/0146167210386733
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.
- Heider, F. (1944). Social perception and phenomenal causality. *Psychological review*, 51, 358.
- Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, 34, 107.
- Jones, E. E. (1990). *Interpersonal perception*. New York, NY: W. H. Freeman.
- Jones, E. E., & Davis, K. E. (1965). From Acts To Dispositions The Attribution Process In Person Perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219–266). New York: Academic Press. doi:10.1016/S0065-2601(08)60107-0
- Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1–24. doi:10.1016/0022-1031(67)90034-0
- Jones, E. E., Worchel, S., Goethals, G. R., & Grumet, J. F. (1971). Prior expectancy and behavioral extremity as determinants of attitude attribution. *Journal of Experimental Social Psychology*, 7, 59–80.
- Kant, I. (1953). *Critique of pure reason*. (N. K. Smith, Trans.). London: Macmillan.
- Kenny, D. A., & Judd, C. M. (2014). Power Anomalies in Testing Mediation. *Psychological Science*, 25, 334–339. doi:10.1177/0956797613502676

- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190–194. doi:10.1111/1467-8284.00419
- Knobe, J. (2014). Free Will and the Scientific Vision. In E. Machery & E. O'Neill (Eds.), *Current Controversies in Experimental Philosophy*. Routledge.
- Koppel, S., Fondacaro, M. R., & Na, C. (in press). Cast into Doubt: Free Will and the Justification for Punishment. *Behavioral Sciences & the Law*.
- Laurent, S. M., Clark, B. A. M., & Schweitzer, K. A. (2015). Why side-effect outcomes do not affect intuitions about intentional actions: properly shifting the focus from intentional outcomes back to intentional actions. *Journal of Personality and Social Psychology*, 108, 18–36. doi:10.1037/pspa0000011
- MacKenzie, M. J., Vohs, K. D., & Baumeister, R. F. (2014). You didn't have to do that: Belief in free will promotes gratitude. *Personality and Social Psychology Bulletin*, 0146167214549322. doi:10.1177/0146167214549322
- Monroe, A. E., Brady, G., & Malle, B. F. (2017). This Isn't the Free Will Worth Looking For: General Free Will Beliefs Do Not Influence Moral Judgments, Agent-Specific Choice Ascriptions Do. *Social Psychological and Personality Science*, 1948550616667616. doi:10.1177/1948550616667616
- Monroe, A. E., Dillon, K. D., Guglielmo, S., & Baumeister, R. F. (2018). It's not what you do, but what everyone else does: On the role of descriptive norms and subjectivism in moral judgment. *Journal of Experimental Social Psychology*, 77, 1–10.
- Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, 27, 100–108. doi:10.1016/j.concog.2014.04.011
- Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, 1, 211–224. doi:10.1007/s13164-009-0010-7
- Monroe, A. E., & Malle, B. F. (2014). Free will without metaphysics. In A. R. Mele (Ed.), *Surrounding free will* (pp. 25–48). New York, NY: Oxford University Press.
- Monroe, A. E., Reeder, G. D., & James, L. (2015). Perceptions of intentionality for goal-related action: Behavioral description matters. *PLoS ONE*, 10, e0119841. doi:10.1371/journal.pone.0119841
- Nahmias, E. (2011). Is Neuroscience the Death of Free Will? *The New York Times*.
- Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, 18, 561–584. doi:10.1080/09515080500264180
- Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction. *Cognition*, 133, 502–516. doi:10.1016/j.cognition.2014.07.009
- Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language*, 19, 473–502. doi:10.1111/j.0268-1064.2004.00269.x
- Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, 41, 663–685.
- Open Science Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. doi:10.1126/science.aac4716
- Overbye, D. (2007, January 2). Free Will: Now You Have It, Now You Don't. *The New York Times*, p. F1.
- 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. doi:10.1080/00223891.2010.528483
- Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20, 30–36. doi:10.1080/10478400902744279
- Reeder, G. D., Monroe, A. E., & Pryor, J. B. (2008). Impressions of Milgram's obedient teachers: Situational cues inform inferences about motives and traits. *Journal of Personality and Social Psychology*, 95, 1–17. doi:10.1037/0022-3514.95.1.1
- Reynolds, W. M. (1982). Development of Reliable and Valid Short Forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 38, 119–125. doi:10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

- Rigoni, D., Kühn, S., Gaudino, G., Sartori, G., & Brass, M. (2012). Reducing self-control by weakening belief in free will. *Consciousness and Cognition*, 21, 1482–1490. doi:10.1016/j.concog.2012.04.004
- Rigoni, D., Kühn, S., Sartori, G., & Brass, M. (2011). Inducing Disbelief in Free Will Alters Brain Correlates of Preconscious Motor Preparation: The Brain Minds Whether We Believe in Free Will or Not. *Psychological Science*, 22, 613–618. doi:10.1177/0956797611405680
- 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. doi:10.1111/j.1468-0017.2010.01393.x
- Stafford, T. (2013). Does non-belief in free will make us better or worse? *The BBC*.
- Stillman, T. F., Baumeister, R. F., & Mele, A. R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology*, 24, 381–394. doi:10.1080/09515089.2011.556607
- Tetlock, P. E., Visser, P. S., Singh, R., Polifroni, M., Scott, A., Elson, S. B., Mazzocco, P., et al. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology*, 43, 195–209. doi:10.1016/j.jesp.2006.02.009
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116, 87–100. doi:10.1016/j.cognition.2010.04.003
- Vonasch, A. J., Baumeister, R. F., & Mele, A. R. (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. *Consciousness and Cognition*, 60, 133–151. doi:10.1016/j.concog.2018.03.002
- Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100, 283–301. doi:10.1016/j.cognition.2005.05.002

Supplementary Materials

Study 1 Vignettes (copied from Clark et al., 2014)

Blameworthy Condition

Nation rocked by 'jailing kids for cash' scandal

At a friend's sleepover more than a year ago, 14-year-old Phillip Swartley pocketed change from unlocked vehicles in the neighborhood to buy chips and soft drinks. The cops caught him.

There was no need for an attorney, said Phillip's mother, Amy Swartley, who thought at most, the judge would slap her son with a fine or community service.

But she was shocked to find her eighth-grader handcuffed and shackled in the courtroom and sentenced to a youth detention center. Then, he was shipped to a boarding school for troubled teens for nine months.

"Yes, my son made a mistake, but I didn't think he was going to be taken away from me," said Swartley, a 41-year-old single mother raising two boys in Wilkes-Barre, Pennsylvania.

The justice system in Luzerne County, Pennsylvania, has fallen prey to corruption. The county has been rocked by a kickback scandal involving two elected judges who essentially jailed kids for cash. Many of the children had appeared before judges without a lawyer.

The nonprofit Juvenile Law Center in Philadelphia said Phillip is one of at least 5,000 children over the past five years who appeared before former Luzerne County Judge Mark Ciavarella. Many of these children had no prior offenses and would normally be shown leniency through being fined or required to do community service. Yet Ciavarella had an ulterior motive- to incarcerate them in order to receive money.

Ciavarella, 58, corruptly and fraudulently "created the potential for an increased number of juvenile offenders to be sent to juvenile detention facilities," federal court documents alleged. Children would be placed in private detention centers, under contract with the court, to increase the head count. In exchange, the two judges would receive kickbacks. Essentially, Ciavarella made millions off of unjustly incarcerating young people.

Morally Neutral Condition

Luzerne County school district starts superintendent search

The Crestwood school district, located in Luzerne County in Pennsylvania, recently announced its intent to look for a new superintendent. The current superintendent, Mark Ciavarella, will retire at the end of this school year.

Ciavarella, who was elected to the position in 1992, has been working on improving the school district for over two decades. Crestwood school district has over 3,000 students, and employs more than 150 teachers. It has two elementary schools, one middle school, and one high school.

Superintendent Ciavarella worked closely with the town mayor, Michael Conahan. The smooth collaboration of the two in the education sector has "made [Crestwood schools] unique," according to the Board of Education curriculum committee chair Susan Samuels.

The city began its search for superintendent last month by hiring search firm PROACT to screen potential candidates. The Board of Education will select the new leader according to a participatory search process involving three open forums — the second of which is taking place this week — and 17 focus groups to hear from the community's stakeholders, including students, parents, teachers and community organizations.

There is so much at stake," said Alex Johnston, a member of the Board of Education. "But at the end of the day we need a leader who can make informed, good decisions." Johnston said that the superintendent's leadership is critical to maintaining the level of achievement that Crestwood public schools have been showing for the past few years.

As the search for the new superintendent gets underway, many of the Board of Education members reflect on the job well done of Superintendent Ciavarella. Board member Alex Johnston says Ciavarella "epitomizes what it means to be an educator."

Study 2 Vignettes (copied from Clark et al., 2014)

Immoral condition

Sam, a special education teacher, wakes up one morning and finds that someone robbed his home while he was sleeping. His window is broken and all of his valuables are missing. After a police investigation, he learns that the robber is unemployed, has two children, and sold all of his belongings on eBay.

Morally neutral condition

Sam, a special education teacher, wakes up one morning and finds that someone rooted through his recycling bin at the end of his driveway while he was sleeping. There is no mess, but all of his aluminum cans are missing. After talking to his neighbors, he learns that the person is unemployed, has two children, and sells the cans to a recycling company.

Study 2 Supplementary Analyses

Study 2a: Analyses for non-moralized free will belief

Matching the result for the full free will subscale, there was no significant effect of the morality manipulation on the non-moralized free will subscale ($\alpha = .76$), $t(399) = -0.253$, $p = .80$, $d = 0.025$, 95% CI [-0.221, 0.170]. Participants belief in free will in the immoral condition ($M = 3.76$, $SD = 0.74$) was statistically identical to the control condition ($M = 3.74$, $SD = 0.83$).

Mediation Analysis. There was no direct effect of condition on free will belief, $b = 0.021$, $se = .079$, 95% CI [-0.134, 0.176], but desire to punish significantly predicted free will belief, $b = 0.081$, $se = .024$, 95% CI [0.034, 0.128]. The overall indirect effect was significant, indirect $b = 0.273$, $se = .080$, 95% CI [0.118, 0.436], but again revealed a suppressor effect. Simultaneously entering condition and desire to punish into the model caused the relationship between condition and free will belief to become significant and negative, direct $b = -0.252$, $se = .112$, 95% CI [-0.473, -0.032].

Study 2b: Analyses for non-moralized free will belief

As in Study 2a, there was no significant effect of the morality manipulation on the non-moralized free will subscale, $t(397) = 0.093$, $p = .926$, $d = 0.009$, 95% CI [-0.187, 0.206]. Participants belief in free will in the immoral condition ($M = 3.64$, $SD = 0.72$) was identical to the control condition ($M = 3.64$, $SD = 0.68$).

Mediation Analysis. There was no direct effect of condition on free will belief, $b = -0.003$, $se = .070$, 95% CI [-0.141, 0.135]. Whereas the desire to punish significantly predicted free will belief, $b = 0.086$, $se = .019$, 95% CI [0.050, 0.123], and the overall indirect effect was significant, indirect $b = 0.187$, $se = .044$, 95% CI [0.107, 0.283], there was also evidence for a suppressor effect. Entering condition and desire to punish into the model simultaneously caused the direct relationship between condition and free will beliefs to become significant, but negative, direct $b = -0.190$, $se = .080$, 95% CI [-0.347, -0.033].

Decomposing Mediation Analyses into zero-order vs. partial correlations

Study 2a. Decomposing the mediation effects into their zero-order and partial correlations showed that the zero-order correlation between condition (1 = neutral behavior; 2 = immoral behavior) and desire to punish was substantial $r(398) = .721$, $p < .001$; whereas the zero-order correlations with the

full free will subscale, $r(398) = .008, p = .87$, and the non-moralized free will subscale, $r(398) = .013, p = .79$, were near zero. However, the partial correlations (controlling for desire to punish), revealed negative correlations between condition and the full free will subscale, $r(397) = -.136, p = .007$, and the non-moralized free will subscale, $r(397) = -.112, p = .025$. That is, the mediation analyses and accompanying partial correlations show that reading about moral violations increased the desire to punish, but did not directly increase free will beliefs. Moreover, the partial correlations show the significant suppressor effect: When controlling for desire to punish, reading about moral violations actually *reduced* people's free will beliefs (opposite Clark et al.'s predicted pattern).

Study 2b. Given the apparent suppressor effects for both the moralized and non-moralized free will beliefs, we again decomposed the effects into their zero-order and partial correlations. As in Study 2a, the zero-order correlation between condition (1 = neutral behavior; 2 = immoral behavior) and desire to punish was substantial, $r(395) = .510, p < .001$; whereas the zero-order correlations with the full free will subscale, $r(395) = .002, p = .97$, and with the non-moralized free will subscale, $r(395) = -.002, p = .97$, were near zero. Again, however, the partial correlations (controlling for desire to punish), revealed negative correlations between condition and the full free will subscale, $r(394) = -.138, p = .006$, and the non-moralized free will subscale, $r(394) = -.119, p = .018$. As in Study 2a, these data demonstrate that reading about a moral violation did not affect free will beliefs relative to control. However, controlling for punishment recommendations, reading about moral violations actually *decreased* people's belief in free will relative to control.

Testing Reverse Mediation Models

Clark et al.'s (2014) core argument is that the desire to punish (measured via recommendations for punishment) motivates people's general belief in free will. The key test for this prediction is a mediation model where punishment judgments are hypothesized to mediate the relationship between condition and free will beliefs. As demonstrated in Studies 2a and 2b there is scant evidence for this hypothesis; however, since regression models rely on correlational methods, it is important to test whether the reverse relationship (free will belief \rightarrow punishment) is also significant, and if so, whether this pathway is stronger or weaker than the punishment \rightarrow free will belief pathway.

We tested this by running the reverse mediation models where measures of people's general free will beliefs (moralized and non-moralized belief in free will) predict punishment decisions (i.e., Immoral behavior \rightarrow Belief in free will \rightarrow Punishment). Although the indirect effects of the models are never significant (indirect $bs < .008$) because, as already demonstrated in Studies 2a and 2b, the condition manipulation always fails to predict free will beliefs ($bs < .003$), the key question is whether the free will belief \rightarrow punishment pathway or the punishment \rightarrow free will belief pathway is stronger.

Table 1 summarizes the direct pathway tests for Study 2a and Study 2b. Across both Studies punishment judgments has a small but consistently significant predictive effect on both moralized and non-moralized free will beliefs. However, examining the reverse pathway shows that the free will belief \rightarrow punishment effect is at least four times larger than the punishment \rightarrow free will belief effect. Examining the 95% confidence interval (95% CI) for the direct effects shows no overlap between the free will \rightarrow punishment and the punishment \rightarrow free will effects demonstrating that the free will \rightarrow punishment effect is statistically significantly stronger than the punishment \rightarrow free will effect, undercutting the core logic of Clark et al.'s theoretical claim.

Table S1. *Comparing direct pathway coefficients for punishment predicting free will beliefs and free will predicting punishment judgments.*

	Direct Effects			
	<i>b</i>	<i>se</i>	<i>p</i>	95% CI
Study 2a				
Punishment → Moralized FW	0.088	0.022	<i>p</i> = .0001	[0.044, 0.132]
Punishment → Non-Moralized FW	0.081	0.024	<i>p</i> = .0008	[0.034, 0.128]
Moralized FW → Punishment	0.429	0.109	<i>p</i> = .0001	[0.216, 0.643]
Non-Moralized FW → Punishment	0.345	0.102	<i>p</i> = .0008	[0.144, 0.546]
Study 2b				
Punishment → Moralized FW	0.093	0.017	<i>p</i> < .0001	[0.060, 0.126]
Punishment → Non-Moralized FW	0.086	0.019	<i>p</i> < .0001	[0.050, 0.123]
Moralized FW → Punishment	0.761	0.139	<i>p</i> < .0001	[0.488, 1.035]
Non-Moralized FW → Punishment	0.591	0.128	<i>p</i> < .0001	[0.339, 0.844]

Study 3 Vignettes

Blameworthy Condition

Andrew is riding the city bus, sitting next to an old lady. A little while later the bus stops, and the lady slowly gets up and exits the bus. As she leaves Andrew notices that she has a fat wallet. Andrew pulls the stop lever and follows her for several blocks. When she turns a corner, Andrew runs up to her, knocks her over, and steals her purse.

Morally Neutral Condition

Andrew is riding the city bus, sitting next to an old lady. A little while later the bus stops, and the lady slowly gets up and exits the bus. As she gets off Andrew slides over to take her place in the seat. He then rides home, making room for people as they come on, and sliding over to take up the entire seat when they get off.

Strange Condition

Andrew is riding the city bus, sitting next to an old lady. A little while later the bus stops, and the lady slowly gets up and exits the bus. As she gets off Andrew decides to yodel an old folk tune to send her on her way. There are only a couple of other people on this bus route and so Andrew decides to ride the bus to the end of the line and cheerfully yodel a folk tune each time a person exits the bus.

Praiseworthy Condition

Andrew is riding the city bus, sitting next to an old lady. A little while later the bus stops, and the lady slowly gets up and exits the bus. A few minutes after the bus starts moving again, Andrew notices that the lady left her purse on the bus. Andrew picks up the purse, pulls the stop lever, and runs five blocks to find the lady, and he gives her back her purse.

Study 3 Stimuli Norming

Prior to conducting Study 3, we conducted a pilot study ($n = 121$) to norm the behaviors for moral valence and strangeness. Participants read one of the four vignettes and indicated: “How much blame or praise does [agent] deserve?” (-5 a lot of blame; 0 neither blame nor praise; +5 a lot of praise) and “How strange or out of the ordinary is [agent]’s behavior?” (1 – not at all strange - 7 Very strange). The pilot confirmed that participants viewed the blameworthy behavior as blameworthy ($M = -4.30$, $SD = 1.95$); they judged the morally neutral ($M = 1.48$, $SD = 1.50$) and the strange ($M = 0.73$, $SD = 2.35$) behavior as neither blameworthy or praiseworthy, and participants judged the praiseworthy behavior as praiseworthy ($M = 4.27$, $SD = 1.23$). Additionally, the data showed that people perceived the strange ($M = 5.47$, $SD = 1.57$) and the blameworthy ($M = 5.07$, $SD = 2.07$) behaviors the most strange, but not significantly different from one another ($p = .353$). Comparatively, people viewed the praiseworthy behavior as less strange ($M = 3.20$, $SD = 1.58$) than the strange or the blameworthy behaviors ($ps < .001$), and the morally neutral behavior ($M = 2.35$, $SD = 1.36$) was judged as the least strange compared to the other behaviors, $ps < .05$.

Study 3 Supplementary Analyses

Exploratory Analyses

In addition to our preregistered hypothesis tests, we conducted two exploratory analyses. First, examining the omnibus effects of condition on free will beliefs showed a significant, though small effect of the moral behavior manipulation on free will beliefs measured with the full, moralized free will scale, $F(3, 796) = 2.76$, $p = .041$, partial $\eta^2 = .010$, 95% CI [0.000, 0.025] and a marginally significant effect on free will beliefs as measured by the reduced, non-moralized free will scale, $F(3, 796) = 2.49$, $p = .059$, partial $\eta^2 = .009$, 95% CI [0.000, 0.024]. There was, however, a significant effect on agent-specific free will judgments, $F(3, 796) = 29.34$, $p < .001$, partial $\eta^2 = .100$, 95% CI [0.062, 0.138]. Importantly, for both measures of free will belief the blameworthy condition did not differ from the praiseworthy or the strange condition ($ps > .15$). Similarly, for agent-specific free will judgments the blameworthy condition did not differ from praiseworthy condition ($p = .13$), though it did differ from the strange condition ($p = .032$).

Second, we tested whether the moral behavior manipulation affected perceptions of a behaviors’ commonness, weirdness, and whether it broke a norm. Univariate ANOVAs showed a significant, effect of the moral behavior manipulation on perceptions of commonness, $F(3, 793) = 38.03$, $p < .001$, partial $\eta^2 = .126$; weirdness, $F(3, 793) = 135.2$, $p < .001$, partial $\eta^2 = .339$; and norm breaking, $F(3, 793) = 179$, $p < .001$, partial $\eta^2 = .401$. Pairwise comparisons showed that all of the behaviors significantly differed from one another with regard to perceived commonness ($ps < .01$), with the exception that praiseworthy behaviors did not differ from control ($p = .47$). Similarly, for perceived weirdness, praiseworthy behaviors did not differ from blameworthy behaviors ($p = .33$), but all of the other behaviors significantly differed from each other ($ps < .001$). Lastly, examining norm breaking showed that all of the behaviors significantly differed from one another ($ps < .001$), with the exception that strange behaviors did not differ from blameworthy behaviors ($p = .45$) (See Figure S1).

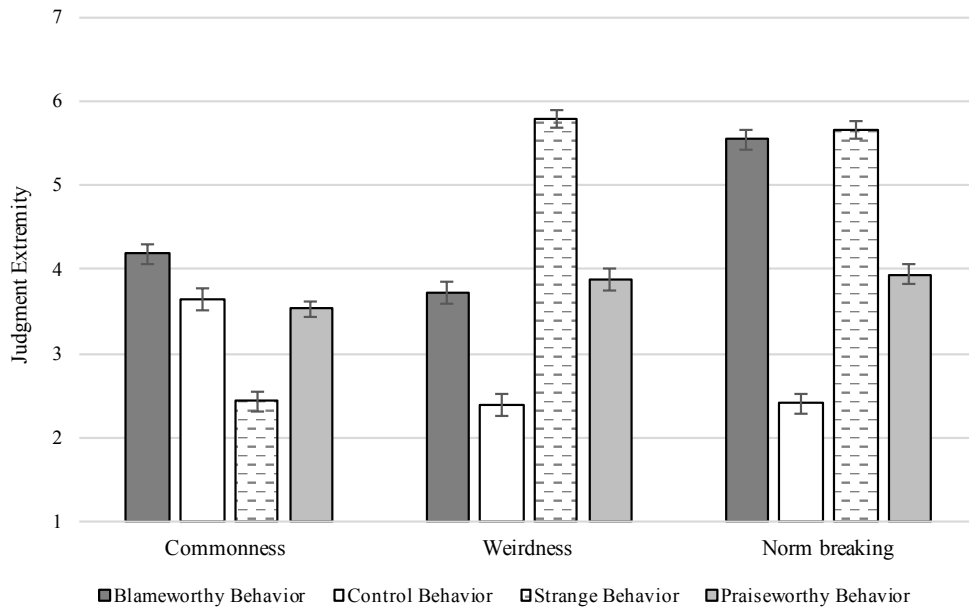


Figure S1. Study 3: Perceived commonness, weirdness, and norm-breaking by moral behavior condition. People viewed strange behaviors least common and most strange, and they judged strange and blameworthy behaviors as most norm breaking. Error bars = ± 1 SE.

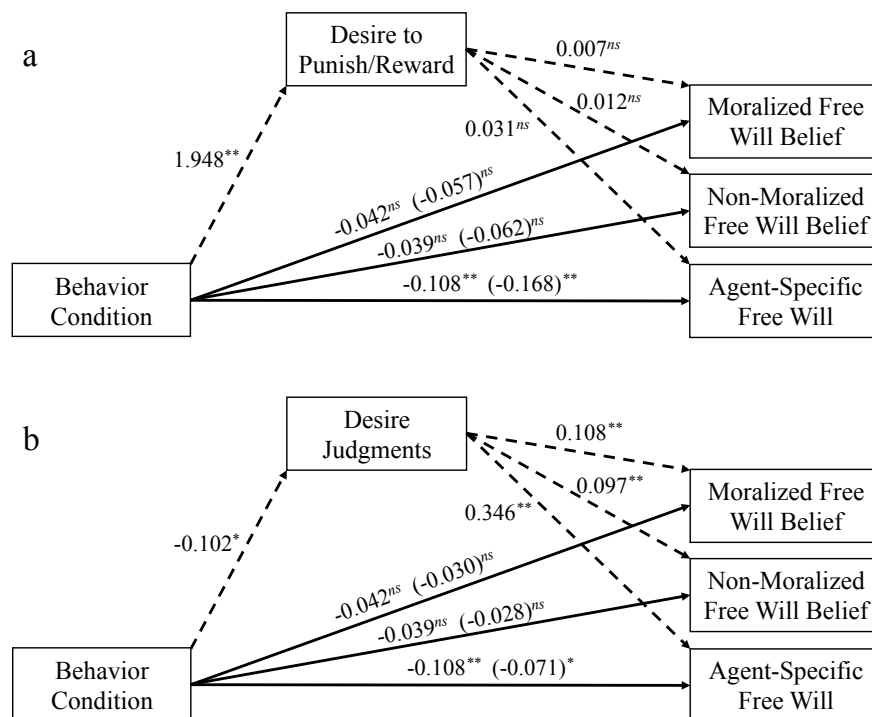


Figure S2. Study 3: The desire to punish failed to predict general free will beliefs and agent-specific free will attributions (panel a). Contrastingly, judgments of agents' desires (panel b) consistently mediated the relationship between the behavior manipulation, and beliefs in free will and agent-specific free will attributions. * $p = .05$, ** $p = .01$,