

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

Andrew E. Monroe  
Appalachian State University

Dominic W. Ysidron  
Ohio University

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. Three replication experiments (Studies 1–2b) attempt to reproduce these findings. Additionally, a novel experiment (Study 3) tests a theoretical challenge derived from attribution theory, which suggests that immoral behaviors do not uniquely influence free will judgments. Instead, our nonviolation model argues that norm deviations of any kind—good, bad, or strange—cause people to attribute more free will to agents. Across replication experiments we found no consistent evidence for the claim that witnessing immoral behavior causes people to increase their general belief in free will. By contrast, we replicated the finding that people attribute more free will to agents who behave immorally compared to a neutral control (Studies 2a and 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, motivated cognition, replication

**Supplemental materials:** <http://dx.doi.org/10.1037/xge0000788.supp>

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, 2016](#); [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, 2015](#)), increases in situational causal attributions ([Genschow, Rigoni, & Brass, 2017](#)), and volition ([Rigoni, Kühn, Sartori, & Brass, 2011](#)).

Recent research, however, suggests a reverse relationship between free will and morality, proposing that moral judgments inform people's belief in free will ([Clark et al., 2014](#)). If correct, this perspective would represent dramatic departure from scholarly and lay conception of free will, which typically hold that free will undergirds human moral behavior and legitimizes legal punishment ([Darwin, 1840](#); [Dennett, 1984](#); [Greene & Cohen, 2004](#); [Nichols, 2011](#)). 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

This article was published Online First June 11, 2020.

 Andrew E. Monroe, Department of Psychology, Appalachian State University; Dominic W. Ysidron, Department of Psychology, Ohio University.

We are indebted to Cory Clark for being willing to share her experimental materials. Without her generosity we would not have been able to complete this project.

Correspondence concerning this article should be addressed to Andrew E. Monroe, Department of Psychology, Appalachian State University, 222 Joyce Lawrence Lane, Boone, NC 28608. E-mail: [monroea1@appstate.edu](mailto:monroea1@appstate.edu)

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 claim would have the potential to significantly disrupt current legal theory, which treats free will as a necessary precursor for justifying punishment (Dennett, 1984; Nichols, 2011). For example, Greene and Cohen (2004) argue that the law implicitly assumes that is permissible to punish offenders' misdeeds only insofar as they had the ability to reasonably "have done otherwise" (a common framing for the expression of free will). 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 heavily on a commonly used measure of free will belief: the Free Will Subscale of the Free Will and Determinism Scale (FAD+; Paulhus & Carey, 2011). This seven-item scale asks people to respond to various statements pertaining to their belief in human agency (e.g., "People have complete control over the decisions they make." "People can overcome any obstacles if they truly want to." "People have complete free will." "Strength of mind can always overcome the body's desires."). However, three of them items appear to confound belief in free will with moral evaluation ("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.").

To date no research we are aware of has examined whether the free will subscale is tapping a unified free will construct or separate agency and morality constructs. Yet, *prima facie* it appears reasonable to hypothesize that assertions that agents "must take full responsibility" or "are always at fault" are punitive judgments, not affirmations of humans' agency or metaphysical freedom. Moreover, this potential duality within the subscale would be problematic for Clark et al. (2014) because the crux of their findings is that observing immoral behaviors on the one hand increases people's belief in free will on the other. If the measure of free will used in their studies is also assessing moral judgments, then it becomes unclear whether their findings reflect people believing more in free will after reading about an immoral behavior or merely condemning that immoral behavior more harshly.

One potential salve to this critique is that the Free Will Subscale is not the only measure of free will included in their studies. Clark et al. demonstrate analogous motivated findings when asking people to make agent-specific free will attributions (e.g., did X exercise her free will)—People are more inclined to say that a person "exercised her free will" after behaving immorally than neutrally. Clark et al. claim that these findings with agent-specific free will attributions insulate their overarching conclusions from criticism. Critically, however, Clark et al. explicitly frame their 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). Immoral behavior influencing agent-specific free will judgments are insufficient to defend such an expansive claim; instead, this claim requires immorality to change people's general belief in free will (as measured by the Free Will Subscale). Yet, it appears plausible that nearly half of the items are contaminated with moral valuation. 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 (a) it does so without positing the additional step of a motivated desire to punish, and (b) 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 people's belief in free will (Studies 1–3) and agent-specific ascrip-

tions of free will (Studies 2a–3). Studies 1 and 2b are close 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 using 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 observed behavior and 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 are publicly available via the OSF ([osf.io/rwe2t](https://osf.io/rwe2t)). We preregistered our sample size, hypotheses, and data analysis plans, and we report all of our measures. Because we wanted to be able to make inferences from the data regardless of the results, we set our desired power at 95% for every study. Our assumed effect size for each of our replication studies was based on the original effects reported in Clark et al. (2014) Study 1 ( $d = .43$ ) and Study 2 ( $d = .47$ ). In all of our replication studies, we assumed slightly more conservative estimates of the effect ( $d = .4$ ), and we oversampled. Further, in Study 3, we address the power and the replicability of our findings by assuming a substantially smaller effect size ( $d = .30$ ) and substantially expanding our sample size ( $n = 800$ ).

## Study 1

### Method

**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 [online 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*). Afterward, participants responded to the 28 item FAD+ (Paulhus & Carey, 2011) which included the free will subscale (see Table 1). 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 6 (“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: (a) 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+). (b) We predicted no significant effect of the moral valence of a target's behavior (immoral vs. neutral) on nonmoralized free will beliefs (measured with the free will subscale of the FAD + omitting the three moralized items).

Both the moralized and the nonmoralized 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 ( $M = 3.71$ ,

Table 1  
The Free Will Subscale of the FAD+ (Paulhus & Carey, 2011)  
With Presumed Moralized Items in Italics

Free will subscale of the FAD+ (Paulhus & Carey, 2011)
1. People have complete control over the decisions they make.
2. <i>People must take full responsibility for any bad choices they make.</i>
3. People can overcome any obstacles if they truly want to.
4. <i>Criminals are totally responsible for the bad things they do.</i>
5. People have complete free will.
6. <i>People are always at fault for their bad behavior.</i>
7. Strength of mind can always overcome the body's desires.

$SD = 0.68$ ),  $t(404) = -0.095$ ,  $p = .925$ ,  $d = 0.009$ , 95% CI  $[-0.204, 0.185]$ . Similarly, nonmoralized 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  $[-0.151, 0.238]$ .

Given the surprising finding that there were no effects of immorality on free will beliefs for either formulation of the free will scale, we conducted two follow-up analyses to test for additional potential differences between the free will measures. First, mixed-model ANOVA with the two free will measures entered as repeated variables to more finely test for potential differences between the free will measures. The analysis revealed no impact of condition,  $F(1, 404) = 0.03$ ,  $p = .85$ ; no condition by measure interaction,  $F(1, 404) = 1.78$ ,  $p = .18$ , but a significant main effect of measure. People reported slightly higher belief in free will when measured with the full free will scale ( $M = 3.72$ ,  $SD = 0.67$ ) compared to the nonmoralized scale ( $M = 3.66$ ,  $SD = 0.73$ ),  $F(1, 404) = 15.48$ ,  $p < .001$ . Second, we conducted a principle component analysis (varimax rotation) to test whether the free will subscale is tapping a unified free will construct or separate agency and morality constructs. The analysis revealed a single factor accounting for 47% of the variance (Eigenvalue = 3.33) with all items loading over .5 on the factor suggesting that people did not respond to the hypothesized “moral” items differently from the other items in the scale.

## Discussion

Our methodological challenge argued that evidence for motivated free will beliefs may be based on a measure that confounds free will belief with moral judgments. Specifically, we predicted that the original findings would replicate when using the full, moralized scale, but that the effects would dissipate once the potentially moralized items were removed from the measure. Instead, in our replication attempt of Clark et al.’s (2014) Study 1, we found no discernable effect of immoral behavior on people’s belief in free will. Exorcising what we believed to be morally tainted items from the free will subscale did not affect any of the results. There simply was no effect present.

Whereas assertions based on 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%. A sensitivity power analysis indicated that we had sufficient power to detect effects as small as  $d = .28$  (with 80% power). 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 focusing on Clark et al.’s (2014) Study 2. Further, these studies 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. Although Study 1 casts doubt on the first premise, Studies 2a and 2b provide further tests of this prediction, and they test the second premise by testing 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.<sup>1</sup> We predict that (a) the moral valence of a target’s behavior (immoral vs. neutral) will increase free will beliefs; (b) the moral valence of a target’s behavior will increase target-specific free will attributions, and (c) that punishment judgments will mediate the relationship between the moral valence manipulation and free will beliefs and target-specific free will attributions.

## Method

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 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 midsize 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 midsize 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 exactly replicated the cover story and stimuli of Clark et al. (2014), Study 2. The dependent variables were the same as those in the original study with the exception that we presented participants with the entire FAD+; whereas Clark et

<sup>1</sup> As in Study 1 we tested the effect of the morality manipulation on the full free will subscale and a reduced nonmoralized version of the scale. However, as in Study 1 the pattern of results for the full scale and the nonmoralized version were identical. We therefore focus on the results from the full free will subscale; results for the nonmoralized scale are reported in the [online supplementary materials](#).



al. (2014) only presented participants with the 7-item free will subscale. 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*). Afterward, 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,  $t(399) = -0.157, p = .875, d = 0.016, 95\% \text{ CI} [-0.211, 0.180]$ . 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,  $t(399) = -3.227, p = .001, d = 0.322, 95\% \text{ CI} [0.125, 0.519]$  (See Figure 1).<sup>2</sup>

Even though there was no 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 (a) full free will subscale and (b) 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 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 nonsignificant 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).

Decomposing these 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 nonmoralized 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 nonmoralized 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.

**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  $b = 0.269, se = .136, 95\% \text{ CI} [0.005, 0.537]$ , and the direct path from condition to free will attributions was no longer significant, direct  $b = 0.085, se = .160, 95\% \text{ CI} [-0.229, 0.399]$  (see Figure 2).

**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,  $t(397) = 0.019, p = .985, d = 0.002, 95\% \text{ CI} [-0.194, 0.198]$ . Unlike Study 2a, the effect of the moral behavior manipulation on agent-specific free will attributions failed to reach the conventional threshold for significance,  $t(397) = -1.875, p = .062, d = 0.188, 95\% \text{ CI} [-0.009, 0.384]$  (See Figure 1).<sup>3</sup>

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,  $b = 2.167, se = .184, 95\% \text{ CI} [1.806, 2.529]$ .

**Free will subscale.** There was no effect of condition on free will belief,  $b = 0.002, se = .064, 95\% \text{ CI} [-0.124, 0.129]$ . Desire to punish significantly predicted free will belief,  $b = 0.093, se = .017, 95\% \text{ CI} [0.060, 0.126]$ , and the overall indirect effect was significant, indirect  $b = 0.201, se = .042, 95\% \text{ CI} [0.126, 0.291]$ . However, as in Study 2a, the analysis revealed that 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  $b = -0.199, se = .072, 95\% \text{ CI} [-0.341, -0.057]$  (See Figure 3).

Given the apparent suppressor effects for both the moralized and nonmoralized 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 nonmoralized 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 nonmoralized free will subscale,  $r(394) = -.119, p = .018$ . As in

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

<sup>3</sup> 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 ( $M = 4.88, SD = 1.82$ ),  $t(395) = -11.80, p < .001, d = 1.180, 95\% \text{ CI} [0.966, 1.393]$ .

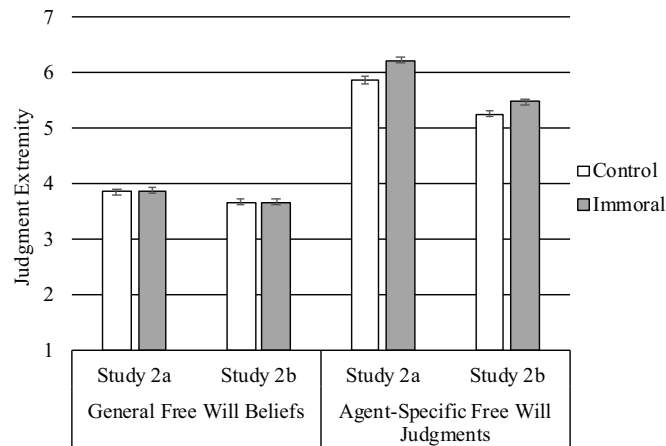


Figure 1. Two studies find no evidence that immoral behaviors motivate general free will beliefs; however, people did attribute more free will to agents who acted immorally versus behaving neutrally. Error bars represent  $\pm 1$  SE.

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.

**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).

## 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 nor 2b replicated Clark et al.'s (2014) findings regarding people's general free will beliefs. This consistent lack of an effect suggests that the effect of immorality on people's 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 suggests 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 consistent 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. Although, the effect in Study 2b fell just

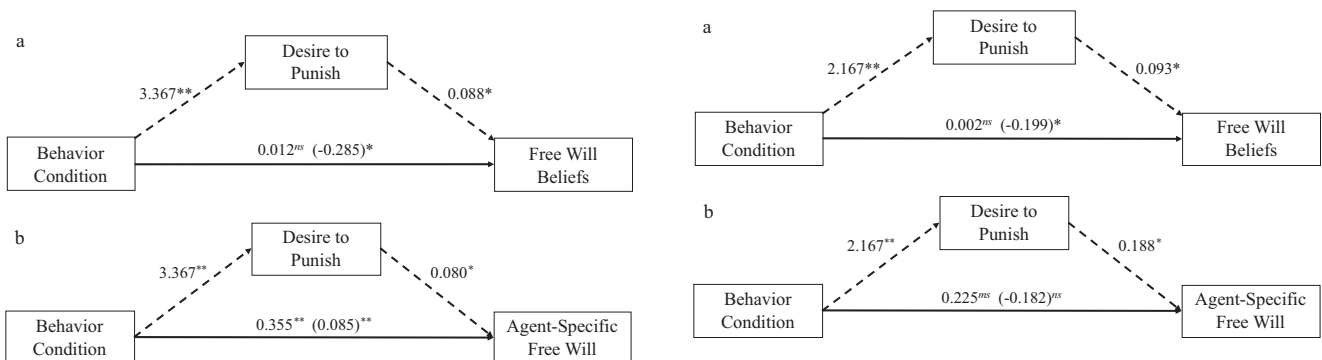


Figure 2. Mediation model demonstrating a suppressor effect. Controlling for punishment judgments, participants report less belief in free will after reading about an immoral behavior compared to a morally neutral control behavior (panel a). Contrastingly, punishment recommendations significantly mediated the relationship between immoral behaviors and agent-specific attributions (panel b). \*  $p < .05$ ; \*\*  $p < .001$ .

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

shy of conventional significance ( $p = .062$ ), it showed the same descriptive pattern. Together, these studies reinforce Clark et al.'s (2014) claim that people attribute more freedom to individuals who behave immorally (compared to a neutral control). However, three replication attempts found no evidence for the broader claim that observing immorality motivated people to increase their general belief in free will itself.

Similarly, regarding the claim 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.

Additionally, the mediation models demonstrated a consistent suppressor effects whereby the previously nonsignificant relationship between condition and free will beliefs became significant and negative. That is, 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 finding that reading about moral violations increases free will beliefs. One possible explanation for this suppressor effect is that there might have been a collinearity problem between the original manipulation (immoral vs. morally neutral) and their punishment measure. The punishment measure might have been 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.

Alternatively, the suppressor effect could be explained via a demand characteristic inherent in the experiment. The assessment of punishment always preceded the free will subscale. Thus, it may be that when people judge a person as responsible for their bad behavior and then respond to free will questions that ask about a person's control and responsibility for their actions they feel an internal pressure for consistency (e.g., "If I thought he deserved to be punished, then he's probably also able to control himself."). Our data are unable to test this possibility, but it would produce results consistent with Clark's original findings as well as the indirect effects in our mediation analyses.

### 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 should be specific to witnessing morally negative behaviors, because only these behaviors would activate a desire to punish. Thus, the motivated free will belief model predicts that blameworthy behaviors should motivate higher free will judgments, relative to control, but other neutral or positive norm violations should not. 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) as well as volition and free will (Monroe, Dillon, Guglielmo, & Baumeister, 2018). Thus, the norm violation model predicts that free will judgments should increase in response to any type of norm violation (i.e., praiseworthy, blameworthy, or strange behavior). To test these predictions, we created four new vignettes that describe an agent acting in either a blameworthy, morally neutral, morally neutral-but-strange, or praiseworthy manner (see the [online supplementary materials](#) for stimuli norming).

### Method

**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 [online 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*). Afterward 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*).<sup>4</sup> Lastly, participants completed the FAD+, 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

<sup>4</sup> We also measured perceptions of an agent's desire to act, commonness, weirdness, and whether a behavior was perceived as breaking a norm as additional measures of the behaviors. We report these data in the [online supplementary materials](#).

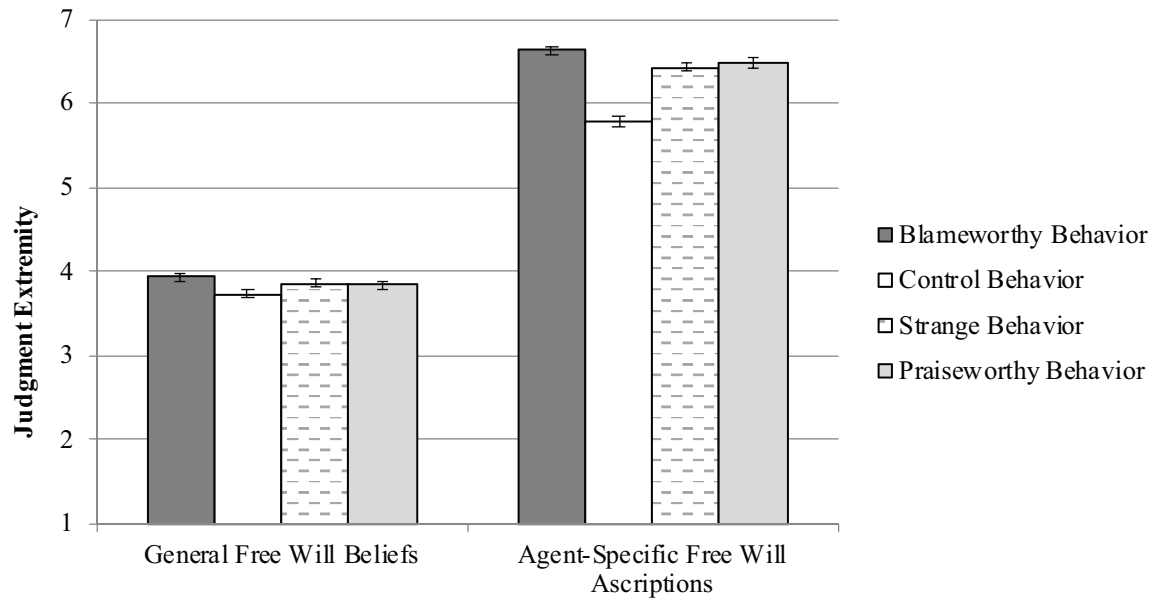


Figure 4. Blameworthy, praiseworthy, and morally neutral-but-strange behaviors all engender stronger free will beliefs and free will judgments relative to control suggesting that norm violation, not immorality, motivates free will judgments. Error bars represent  $\pm 1$  SE.

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.<sup>5</sup> However, the two models starkly diverged for agent-specific free will ascriptions. Although 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]. 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.

Additionally, simple effects revealed mixed evidence for motivated free will belief. Although, relative to control, people attributed the most free will to agents in the blameworthy condition ( $p < .001$ ,  $d = 0.82$ ), this condition did not differ from the praiseworthy condition ( $p = .13$ ,  $d = 0.17$ ), and differences from the strange condition were significant but 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 ( $p < .001$ ,  $d = 0.59$ ; See Figure 4).

## Discussion

Study 3 revealed two key findings. First, consistent with Studies 1 and 2, we demonstrated that the moral valence of agents' behaviors had 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.

Second, 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. These findings suggest a broader explanatory mechanism than free will being motivated by a desire to punish. The data suggest that any behavior

<sup>5</sup> Examining nonmoralized 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].



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.” (p. 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 clearly replicated the 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 nonmoral 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 a strange act were judged as having more free will than an 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 behaviors.

Together these findings argue for a nonmoral 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 do not, 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. Both the replication studies (Studies 1–2b), and the novel experiment (Study 3) found no evidence that witnessing immoral behavior causes people to increase their general belief in free will. However, we did consistently replicate the finding that people ascribe more free will to agents who behave immorally (Studies 2a and 3); however, Study 3 suggests that this effect is best 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.<sup>6</sup>

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 at-

<sup>6</sup> 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 a 1–5 scale on how seriously they filled out the survey. Including these 11 participants reduced the effect below conventional significance ( $p = .077$ ,  $d = 0.24$ , 95% CI [−0.025, 0.500]).

tempts—then it is no longer parsimonious to favor the motivated free will interpretation among the range of possible explanations.

Finally, since we completed our studies, Clark and colleagues conducted a series of new experiments addressing our replication attempts and theoretical critiques (Clark, Winegard, & Shariff, 2020). Across an impressive array of 14 well-powered studies they demonstrate three key findings that reinforce several of our conclusions, challenge others, and suggest a potential accommodation between our seemingly competing theories. First, across studies they show consistent evidence that immoral behavior substantially increases agent-specific free will attributions relative to a neutral control. Additionally, their new data reveal that bad behaviors always produce the largest differences from a neutral control. Overall, however, as in our Study 3, they found that free will attributions following bad behaviors did not differ from those following good behaviors. Second, they demonstrate that the impact of immorality on free will attributions holds even when immoral behaviors are less norm-violating than control behaviors. Third they, show that the impact of immorality on free will beliefs is exceptionally tenuous. The effect was present in only 50% of their studies and was partially dependent on how free will belief was measured.

Together, these new data reinforce some of our findings and require accommodation and nuance for some of our conclusions. They reinforce our work by further demonstrating the robust impact of immorality on agent-specific free will attributions. This effect is clear in the original studies by Clark, in our replication studies, and in Clark's new data. Thus, going forward, this appears a best candidate for studying motivated free will effects. Additionally, the new data reinforce our challenge to the claim that immorality motivates people's general belief in free will. Across our replication and novel experiments, we found little-to-no evidence for this effect. In Clark's new data, the prevalence of the effect is at chance, when present the effect is generally small, and the presence of the effect appears to be dependent on how free will beliefs are assessed. Thus, we would argue that, though we cannot definitively rule this effect out, its small and inconsistent nature renders it an unlikely effect at best. Lastly, the new data require some accommodation on our part. The finding that immoral behavior (controlling for its degree of counternormativity) increases free will attributions coupled with the finding that bad and good behaviors produced statistically indistinguishable increases in free will judgments suggests that the motivated free will and the norm-violation theories are not mutually exclusive. Although the norm violation model may explain a large swath of what motivates people's free will attributions, immorality accounts for unique variance in free will judgments that our norm-violation model cannot.

## 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, 2019). 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, 2018; 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., 2014, 2017; Reeder, Monroe, & Pryor, 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 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 affect 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. <http://dx.doi.org/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. <http://dx.doi.org/10.1016/j.jesp.2012.08.015>
- 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. <http://dx.doi.org/10.1177/0146167214563673>
- Aristotle. (1985). *Nicomachean ethics* (T. Irwin, Trans.). Indianapolis, IN: Hackett.
- Baumeister, R. F., Masicampo, E. J., & Dwall, 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. <http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/10.1037/a0035880>

- Clark, C., Winegard, B., & Shariff, A. (2020). *Motivated free will belief: The theory, new (preregistered) studies, and three meta-analyses*. Manuscript in preparation.
- Crescioni, A. W., Baumeister, R. F., Ainsworth, S. E., Ent, M., & Lambert, N. M. (2016). Subjective correlates and consequences of belief in free will. *Philosophical Psychology*, 29, 41–63. <http://dx.doi.org/10.1080/09515089.2014.996285>
- Crone, D., & Levy, N. L. (2019). Are free will believers nicer people? (Four studies suggest not). *Social Psychological and Personality Science*, 10, 612–619. <http://dx.doi.org/10.1177/1948550618780732>
- Darwin, C. R. (1840). *Old and useless notes about the moral sense and some metaphysical points*. (P. H. Barrett, Trans.). Retrieved from <http://darwin-online.org.uk/content/frameset?viewtype=side&itemID=CUL-DAR91.4-55&pageseq=1>
- Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.
- 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, pp. 307–338). San Diego, CA: 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 of the United States of America*, 114, 10071–10076. <http://dx.doi.org/10.1073/pnas.1701916114>
- Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 359, 1775–1785. <http://dx.doi.org/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. <http://dx.doi.org/10.1177/0146167210386733>
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.
- Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, 51, 358–374. <http://dx.doi.org/10.1037/h0055425>
- Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, 34, 107–117. <http://dx.doi.org/10.1037/0003-066X.34.2.107>
- Jones, E. E. (1990). *Interpersonal perception*. New York, NY: Freeman.
- Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process. In I. L. Berkowitz (Ed.), *Advances in experimental social psychology: Vol. 2. Person perception* (pp. 219–266). New York, NY: Academic Press. [http://dx.doi.org/10.1016/S0065-2601\(08\)60107-0](http://dx.doi.org/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. [http://dx.doi.org/10.1016/0022-1031\(67\)90034-0](http://dx.doi.org/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. [http://dx.doi.org/10.1016/0022-1031\(71\)90055-2](http://dx.doi.org/10.1016/0022-1031(71)90055-2)
- Kant, I. (1953). *Critique of pure reason* (N. K. Smith, Trans.). London, UK: Macmillan.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190–194. <http://dx.doi.org/10.1093/analysis/63.3.190>
- Knobe, J. (2014). Free will and the scientific vision. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 69–85). New York, NY: Routledge. <http://dx.doi.org/10.4324/9780203122884-5>
- Koppel, S., Fondacaro, M., & Na, C. (2018). Cast into doubt: Free will and the justification for punishment. *Behavioral Sciences & the Law*, 36, 490–505. <http://dx.doi.org/10.1002/bsl.2356>
- 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. <http://dx.doi.org/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*, 40, 1423–1434. <http://dx.doi.org/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*, 8, 191–199. <http://dx.doi.org/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. <http://dx.doi.org/10.1016/j.jesp.2018.03.010>
- 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. <http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/10.1093/acprof:oso/9780199333950.003.0003>
- Nahmias, E. (2011, November 13). Is neuroscience the death of free will? *The New York Times*. Retrieved from <https://opinionator.blogs.nytimes.com/2011/11/13/is-neuroscience-the-death-of-free-will/>
- Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, 18, 561–584. <http://dx.doi.org/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. <http://dx.doi.org/10.1016/j.cognition.2014.07.009>
- Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language*, 19, 473–502. <http://dx.doi.org/10.1111/j.0268-1064.2004.00269.x>
- Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, 331, 1401–1403. <http://dx.doi.org/10.1126/science.1192931>
- Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41, 663–685. <http://dx.doi.org/10.1111/j.1468-0068.2007.00666.x>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <http://dx.doi.org/10.1126/science.aac4716>
- Overbye, D. (2007, January 2). Free will: Now you have it, now you don't. *The New York Times*. Retrieved from <https://www.nytimes.com/2007/01/02/science/02free.html>
- 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. <http://dx.doi.org/10.1080/00223891.2010.528483>

- Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20, 30–36. <http://dx.doi.org/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. <http://dx.doi.org/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. [http://dx.doi.org/10.1002/1097-4679\(198201\)38:1<119::AID-JCLP2270380118>3.0.CO;2-I](http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/10.1111/j.1468-0017.2010.01393.x>
- Stafford, T. (2013, September 24). Does non-belief in free will make us better or worse? *The BBC*. Retrieved from <https://www.bbc.com/future/article/20130924-how-belief-in-free-will-shapes-us>
- 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. <http://dx.doi.org/10.1080/09515089.2011.556607>
- Tetlock, P. E., Visser, P. S., Singh, R., Polifroni, M., Scott, A., Elson, S. B., . . . Rescobar, P. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology*, 43, 195–209. <http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/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. <http://dx.doi.org/10.1016/j.cognition.2005.05.002>

Received August 5, 2019

Revision received February 3, 2020

Accepted April 2, 2020 ■