Consciousness and Cognition 51 (2017) 193–211

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Making punishment palatable: Belief in free will alleviates punitive distress☆

Cory J. Clark[a,b,d,*], Roy F. Baumeister[a,c], Peter H. Ditto[d]

[a] Florida State University, Department of Psychology, 1107 West Call Street, Tallahassee, FL 32306-4301, USA
[b] University at Buffalo, The State University of New York, Department of Psychology, Park Hall Room 204, Buffalo, NY 14260-4110, USA
[c] The University of Queensland, School of Psychology, Sir Fred Schonell Drive, St Lucia, QLD 4072, Australia
[d] University of California, Irvine, Department of Psychology and Social Behavior, 4201 Social and Behavioral Sciences Gateway, Irvine, CA 92697-7085, USA

ARTICLE INFO

Keywords:
Free will
Punishment
Morality
Motivated reasoning
Anxiety

ABSTRACT

Punishing wrongdoers is beneficial for group functioning, but can harm individual well-being. Building on research demonstrating that punitive motives underlie free will beliefs, we propose that free will beliefs help justify punitive impulses, thus alleviating the associated distress. In Study 1, trait-level punitiveness predicted heightened levels of anxiety only for free will skeptics. Study 2 found that higher state-level incarceration rates predicted higher mental health issue rates, only in states with citizens relatively skeptical about free will. In Study 3, participants who punished an unfair partner experienced greater distress than non-punishers, only when their partner did not have free choice. Studies 4 and 5 confirmed experimentally that punitive desires led to greater anxiety only when free will beliefs were undermined by an anti-free will argument. These results suggest that believing in free will permits holding immoral actors morally responsible, thus justifying punishment with diminished negative psychological consequences for punishers.

1. Introduction

Back during much of the 20th century, when even the most conscientious parents recognized a duty to spank their children for misbehavior, they would sometimes tell the child beforehand, “This hurts me more than it hurts you.” This assertion was an inviting target for comedy writers, but it captures a dilemma that lies at the heart of not only parenting, but socialization, job training, law enforcement, war, revenge, and many other situations that confront human beings with the prospect of punishing one another: Administering punishment is often aversive. People who must punish others may therefore seek ways of making it more palatable.

The prospect of punishing another adult who misbehaves evokes two contrary impulses. Both are likely deeply rooted in evolution and human nature. The first is a basic reluctance to harm another human being (e.g., Cushman, Gray, Gaffey, & Mendes, 2012). The second is a strong inclination to punish those who pose harm to the self or social group (Fehr & Gächter, 2002). The present investigation was inspired by the assertion that exposure to the harmful actions of others, and subsequent motives to punish such actions, underlie the belief in human free will (Clark et al., 2014). We propose that believing in free will is instrumental and pragmatically helpful in enabling people to administer punishment without suffering the remorse that normally attends harming another person.

☆ The work of Cory J. Clark was supported in part by a grant from The Charles Koch Foundation.
* Corresponding author at: Florida State University, Department of Psychology, 1107 West Call Street, Tallahassee, FL 32306-4301, USA.
E-mail address: cclark3@fsu.edu (C.J. Clark).

http://dx.doi.org/10.1016/j.concog.2017.03.010
Received 8 October 2016; Received in revised form 19 January 2017; Accepted 27 March 2017
1053-8100/ Published by Elsevier Inc.

1.1. Harm aversion

A universal foundation of morality is that harming others is wrong (e.g., Graham et al., 2013; Gray, Schein, & Ward, 2014). Besides the risks of social disapproval and punishment, harming others, or even thinking about harm to another, produces direct negative physiological and emotional consequences (e.g., Baumeister, Stillwell, & Heatherton, 1994; Cushman et al., 2012; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Even Nazis explicitly strategized on how to overcome "animal pity" in their killing, the innate aversion to witnessing human suffering (Arendt, 1963). One of the most compelling works on the struggles of killing was by Browning (1993), with the cogent title *Ordinary Men*. Browning, a historian, wrote about a troop of middle-aged German policemen who were sent to duty in occupied Poland and then abruptly one morning were assigned to shoot all the Jews in a small town. These men experienced severe stress, including anxiety, nightmares, disobedience, and gastrointestinal disturbances. The policemen often struggled with the problem of "shooting past." The civilian would lay face down on the ground while the policeman aimed a gun directly at the back of his or her head, pulled the trigger—and missed. At the last minute the policeman's body involuntarily jerked the gun slightly so as to avoid killing another human being at point-blank range.

The reluctance to harm others is widespread, even in circumstances that would justify or even positively value violent aggression. The standard example would be the battlefield, in which soldiers seek to kill their opponents. All values support them doing so: They are doing their soldiers' duty, protecting their country from its enemies, and crucially, preserving their own lives by eliminating people who want to kill them. Yet soldiers experience considerable difficulty in bringing themselves to kill the enemy (Grossman, 1996). Why are soldiers reluctant to do what duty and even self-preservation dictate they should do? George Orwell (1943), who served in the Spanish Civil War, writes that the difficulty stems from harming a fellow *human*. He describes an opportunity to snipe an enemy soldier, who emerged from the trenches in plain sight half-dressed, holding up his trousers, "I refrained from shooting *him*… I did not shoot partly because of that detail about the trousers… a man who is holding up his trousers isn't a 'Fascist', he is visibly a fellow-creature, similar to yourself, and you don't feel like shooting at him." Apparently soldiers too are reluctant to harm other human beings, even their deadly enemies.

1.1.1. Punishment as harm

Punishment is a particular type of harm directed toward others who have done harm themselves, and thus, some might assume that such harm would be free of aversive feelings. Consistent with that view, research on the neural bases of punishment has demonstrated that the anticipation of punishing a norm violator sometimes activates a region of the brain associated with reward and pleasure (de Quervain et al., 2004). However, although the notion that "revenge is sweet" appears reflected in the beliefs of ordinary people, it may be somewhat misguided. In three studies, Carlsmith, Wilson, and Gilbert (2008) demonstrated the "paradoxical consequences of revenge": participants anticipated that punishing free riders would make them feel better, but in actuality, participants given the opportunity to punish felt worse than those not given the opportunity. Furthermore, participants anticipated feeling equally satisfied when someone else delivered punishment to free riders as when they delivered the punishment themselves, but personally delivering the punishment was more affectively costly. In a similar vein, Bushman, Baumeister, and Phillips (2001) found that American college students would aggress against someone who had insulted them—but only insofar as they expected to feel better afterward. A bogus mood-freezing pill (that ostensibly rendered emotional states temporarily impervious to change) eliminated the link between anger and aggression. Thus, angry people aggress because they expect to feel better afterward, but in actuality, aggressors experienced more hostile and negative affect after aggressing.

The damaging effects of punitiveness are also reflected by the heightened prevalence of mental health issues among those called upon to deliver punishment on society's behalf (i.e., corrections professionals; Spinaris, Denhof, & Kellaway, 2012). The practice of mixing guns loaded with blank cartridges among a firing squad's rifles caters to the wish to sustain the possibility that oneself did not actually kill the target, again suggesting that punishing is aversive. Even making the decision to punish without carrying out the punishment oneself can take an emotional, psychological, and even physical toll. Many jurors who have served on capital trials report experiencing mental and emotional upset for weeks or even months after the trial, causing relationship problems, difficulty sleeping, and physical illness (Antonio, 2006). Furthermore, capital trial jurors whose verdicts resulted in a death sentence experienced greater symptoms of PTSD than capital trial jurors whose verdicts did not render a death sentence (Cusack, 1999). It appears it can be difficult to punish or even vote to punish people, even those who commit the most heinous acts of violence.

In fact, a wide range of deleterious effects is associated with being punitive as opposed to forgiving. Being punitive toward others is associated with higher depression (Maltby, Macaskill, & Day, 2001); vengefulness is associated with greater rumination, higher negative affect, and lower life satisfaction (McCullough, Bellah, Kilpatrick, & Johnson, 2001); venting anger actually leads to greater anger and aggression (Bushman, 2002); and anger expression has been linked to higher negative affect, anxiety, and lower quality of life (Phillips, Henry, Hosie, & Milne, 2005). Similar work has linked unforgiveness with higher depression and stress; worse subjective, psychological and physical well-being; lower satisfaction with life; and more negative and reduced positive moods (e.g., Bono, McCullough, & Root, 2007; Lawler-Row, Karremans, Scott, Edlis-Matityahou, & Edwards, 2008; Lawler-Row & Piferi, 2006).

Taken together, the evidence indicates that people in general have a broad reluctance to administer harm to other people and suffer a variety of psychological and physiological consequences as a result. A state of intense anger, combined with the expectation that aggressing may feel good, can overcome this to some degree, but the expectation that harming others will feel good may often be proven wrong when the moment arrives.

1.2. The benefits of punishment

Like other animals, humans have evolved to learn from punishment. This means that humans can use punishment as an effective tool for teaching fellow humans to cooperate (Cushman, 2013). When given the opportunity, people often behave selfishly by contributing less than their share (e.g., Karau & Williams, 1993; Kerr & Bruun, 1983; Latané, Williams, & Harkins, 1979). Such selfish tendencies carry great weight in social interactions, as the selfish behavior of one individual can lead cooperators to defect, leading to collective disaster for the whole group (Kerr, 1983; Orbell & Dawes, 1981). In response to such selfish behavior, humans demonstrate strong inclinations to punish. Classic studies by Fehr and Gächter (2002) showed that many people engage in so-called altruistic punishment, even in ad hoc and transient laboratory groups. Participants who played a resource dilemma game with selfish confederates would administer punishment to the confederate, even at cost to themselves. This posed a challenge to some classic economic theories, by which people maximize self-interest. But apparently people come to feel they have a stake in maintaining group cohesion, and are willing to assume a cost to themselves in order to punish those who undermine the collective benefit of the group. This tendency to engage in costly punishment is prevalent across cultures (e.g., Henrich et al., 2005, 2006) and carries great benefit for living in social groups (e.g., Clutton-Brock & Parker, 1995; Henrich et al., 2010). For example, more punitive societies demonstrate higher levels of altruism (Henrich et al., 2006), and the possibility of punishment leads people to behave less selfishly (e.g., Fehr, Gächter, & Kirchsteiger, 1997).

Although punishment effectively deters antisocial behavior (and people express support for utilitarian forms of punitiveness), the actual motivations underlying punishment appear predominantly retributive (e.g., Carlsmith, Darley, & Robinson, 2002). From an evolutionary standpoint, this innate "taste" for retribution deters selfish behavior without the cost of cognitive elaboration on the likely long-term benefits of each opportunity to punish (Cushman, 2013). In other words, although punishment may certainly be justified on the grounds of deterrent benefits, this does not appear to be the proximate reason why humans punish one another. Rather, humans *want* to punish others for their harmful behaviors.

All modern countries have institutions, especially police and legal systems, to punish those who break the rules. However, even outside of these institutions where punishment is regarded as legally and socially acceptable, people are compelled to punish others who cause harm. The present work focuses on these individuals. Ordinary people recognize the value and even obligation to punish those who behave badly, much like a parent feels an obligation to punish their child. Yet harming a fellow human being can be psychologically challenging. Thus, punishment presents a dilemma that must be overcome.

1.3. Making punishment palatable

That punishment can have negative emotional consequences leads one to wonder how people are so able and willing to punish. An inherent taste for retribution carries great benefit for living in social groups, and yet our punitive impulses can also be distressing and self-destructive. We hypothesized that belief in free will facilitates the ability to punish, alleviating the distress that would typically arise from harming a fellow human being by establishing that the person deserved punishment.

Despite scientific challenges to the existence of human free will (e.g., Bear & Bloom, 2016; Libet, 1985; Soon, Brass, Heinze, & Haynes, 2008; Wegner, 2002; Wegner, Sparrow, & Winerman, 2004); and despite a persistent lack of consensus among philosophers, psychologists, and neuroscientists regarding free will; the vast majority of laypeople believe in the human capacity for free action (e.g., Nahmias, Morris, Nadelhoffer, & Turner, 2005; Nichols, 2004; Sarkissian et al., 2010). Why is there such widespread public agreement on one of the most contentious philosophical debates of all time? Previous research has demonstrated a number of factors, such as the powerful subjective experience of freedom or observations about causality (e.g., Nichols, 2004; Wegner, 2002, 2003), but one is a fundamental desire to punish wrongful behavior (Clark et al., 2014). In five experiments, Clark et al. (2014) demonstrated that people increased their belief in free will after exposure to immoral behavior and the subsequent desire to punish. For example, students who believed a fellow classmate had cheated on a recent midterm exam reported higher belief in free will than students not informed of a cheating incident, and this was mediated by their endorsement of harsher punishment for cheaters. In another study, it was found that countries with higher rates of crime and homicide (and thus more exposure to harmful behavior) had stronger country-level free will beliefs. These results suggest that one explanation for the strength and prevalence of free will beliefs is an underlying desire to blame and punish others for their misbehavior, supporting a proposition introduced by Nietzsche (1889/1954).

Relevant here is the idea that "blame requires warrant" (Malle, Guglielmo, & Monroe, 2014). Blame and punishment are harmful to those blamed and punished, and therefore must be justified. In fact, people will punish those who punish others, and they do so most harshly when the initial punishment was undeserved (Cinyabuguma, Page, & Putterman, 2006). It is evident from common sense that people should only be blamed for behaviors for which they were responsible, that is, behaviors that were freely chosen. This principle is reflected in both philosophical and psychological theory (Darley & Shultz, 1990; Roskies & Malle, 2013), and in the judgments of laypeople (e.g., Nichols & Knobe, 2007), even children as young as 7-years old (Darley & Zanna, 1982). Moreover, people endorse more retributive forms of punishment for transgressions within an agent's control (e.g., Shariff et al., 2014; Weiner, Graham, & Reyna, 1997). As a practical example, consider the lenient treatment of crimes committed by those with limited control (e.g., children, the mentally ill). A general disbelief in free will would extend such leniency to all criminals.

That people bolster their belief in free will when exposed to harmful actions (Clark et al., 2014) suggests that the motivation to blame influences the very judgments by which blame is warranted. Moral judgments are often susceptible to these sorts of emotional influences, and people feel compelled to produce rational explanations in order to justify those judgments (e.g., Alicke, 2000; Clark, Chen, & Ditto, 2015; Haidt, 2001). We propose that people's tendency to bolster their belief in free will after experiencing punitive

motives serves the function of retroactively fulfilling a requirement for moral responsibility (namely, control), which in turn, helps justify their punitive impulses. In other words, believing in free will enables people to perceive miscreants as responsible for their misdeeds, so punishment is justified.

1.4. Sweet revenge and palatable punishing

It is easy to think of instances when people seem to derive pleasure from others being punished, for example, the celebrations in Libya following the brutal public death of Muammar Gaddafi and American celebrations following the killing of Osama Bin Laden. These examples render implausible any assertion that punishment is *always* distressing or *never* satisfying to punishers or observers of punishment. Our central hypothesis is that framing the punishment as appropriately justified reduces the negative and increases the positive feelings about punishment.

In fact, some work has suggested that it is specifically in cases when punishment is justified that it can have positive affective consequences. Work by Gollwitzer and colleagues (Funk, McGreer, & Gollwitzer, 2014; Gollwitzer & Denzler, 2009; Gollwitzer, Meder, & Schmitt, 2011) demonstrated that people can feel satisfied when unfair others are punished, specifically, when the punished other acknowledged that they deserved the punishment brought upon them and/or indicated that they would behave better in future situations. The acknowledgment that the punishment was deserved and the demonstration that the punishment was effectual may serve as justifications for the punisher's actions, thus leading to more positive affective outcomes.

Thus, we propose that people approve of well-deserved punishment in the abstract but are often reluctant to advocate or administer punishment to actual persons—and so by elevating their belief that the wrongdoer freely and deliberately chose to misbehave, they can reduce their negative feelings and possibly increase positive ones. In this connection, it is noteworthy that belief in free will is the default position in most societies worldwide. That is, people assume others have free will (e.g., Nichols, 2004; Sarkissian et al., 2010), and thus punishment ought to be justified in many, if not most cases (and so punishers should not feel guilty). The present research aimed to show that in the absence of free will, punishment is particularly distressing. Thus, we hypothesized that ascribing free will to immoral actors is one means by which people justify the punishment of harmful actors, therefore alleviating the distress typically experienced when fellow human beings are harmed. To be clear, we are not making normative claims about the ethics of punishment or whether free will beliefs ought to be bolstered or suppressed. Though we described the tendency for people to increase their belief in free will after experiencing punitive motives as potentially useful for reducing punitive distress, the ethics of punishment is a complicated issue that goes far beyond the scope of the present research.

1.5. The present research

Because free will is considered a prerequisite for traditional moral responsibility (Nichols & Knobe, 2007), and because people bolster their belief in free will when motivated to hold others morally responsible, we propose that belief in free will reduces punitive distress by justifying punitive responses. The central hypothesis was that higher belief in free will would be linked to less distress over punishing others. Therefore, we predicted that reduced beliefs in free will would exacerbate the influence of punitiveness on psychological distress. Though punitiveness has been shown to have a wide range of negative consequences for health and well-being, in the present research, we focus on state-distress (Studies 3–4), state-anxiety (Studies 3 and 5), trait-anxiety (Study 1), and general mental health (Study 2). We tested our hypothesis across five studies by determining whether free will beliefs moderate the relationship between punitiveness and distress at three different levels of analysis: individual differences (Study 1), societal-level (Study 2), and experimental (Studies 3–5).

2. Study 1

Before proceeding to the main tests of our hypothesis (the experiments in Studies 3–5), we report two correlational studies intended to show that the effect is not confined to artificial laboratory conditions. They may be considered validation studies.

Study 1 sought initial support for the hypothesis that free will beliefs moderate the relationship between punitiveness and psychological well-being by examining these relationships on the level of individual differences. The prediction was that for those relatively skeptical about free will, high punitiveness would be linked to high anxiety—but that there would be no relationship between punitiveness and anxiety for those who believe strongly in free will.

2.1. Method

All individuals who completed each of three separate scales (described next) on yourmorals.org (a public website on which visitors self-select to complete one or more of several surveys in exchange for response feedback) were included as participants (n = 470; M_{age} = 40.76; 179 female). The Free Will and Determinism scale (FAD; Paulhus & Margesson, 1994; α = 0.79) contains a subscale measuring free will belief on seven items (e.g., "People have complete control over the decisions they make."), each rated on a 5-point scale from "Totally disagree" to "Totally agree". The Comprehensive Justice Scale (Gromet, Haidt, & Darley, 2013) contains a traditional justice subscale measuring punitiveness on six items (e.g., "An eye for an eye is the correct philosophy behind punishing offenders."), each rated on a 7-point scale from "Strongly disagree" to "Strongly agree", α = 0.80. An anxiety measure was adapted from the Gallup poll, on which participants rated how distressed they were during the past 7 days by three anxiety symptoms (e.g., "feeling fearful"), each rated on a 5-point scale from "Not at all" to "Extremely", α = 0.83. Demographic information was collected at

Table 1
Anxiety regressed on punitiveness, free will belief, and interaction with and without controls in Study 1.

	F	R^2	*b*	*SE*	*t*	*p*
Anxiety (Model)	8.71	0.05				< 0.001
Punitiveness			0.10	0.03	3.25	0.001
Free Will Belief			−0.12	0.05	−2.51	0.012
Punitiveness × Free Will Belief			**−0.16**	**0.03**	**−4.63**	**< 0.001**
Anxiety (Model)	6.71	0.10				< 0.001
Age			0.00	0.00	−1.90	0.058
Sex			−0.08	0.07	−1.23	0.219
Education			−0.04	0.01	−2.95	0.003
Religiosity			0.03	0.02	1.68	0.094
Political Ideology			−0.04	0.02	−1.87	0.062
Punitiveness			0.12	0.03	3.66	< 0.001
Free Will Belief			−0.12	0.05	−2.44	0.015
Punitiveness × Free Will Belief			**−0.16**	**0.03**	**−4.62**	**< 0.001**

registration including age, sex, level of education (on a 9-point scale from “some high school” to “completed graduate or professional degree”), religious attendance (on a 5-point scale from “never” to “one or more times a week”), and general political ideology (on a 7-point scale from “very liberal” to “very conservative”). All predictor variables were mean-centered.

2.2. Results

Anxiety was regressed on free will beliefs, punitiveness, and the interaction controlling for age, sex, education, religiosity, and political ideology. As can be seen in Table 1, greater punitiveness predicted higher anxiety, $\beta = 0.20$, $t = 3.66$, $p < 0.001$, 95% CI [0.053, 0.176], with a small effect size (partial $r = 0.17$, semipartial $r = 0.16$). These results are consistent with prior work demonstrating that highly punitive individuals generally have worse well-being than others (e.g., McCullough et al., 2001). Consistent with prior work demonstrating that high free will beliefs are associated with life satisfaction (Baumeister & Brewer, 2012), higher free will beliefs predicted lower anxiety, $\beta = -0.13$, $t = -2.44$, $p = 0.015$, 95% CI [−0.211, −0.023], with a small effect size (partial $r = -0.11$, semipartial $r = -0.11$).

Consistent with our hypothesis, we found a significant punitiveness × free will belief interaction, $\beta = -0.22$, $t = -4.62$, $p < 0.001$, 95% CI [−0.223, −0.090], with a small to medium effect size (partial $r = -0.21$, semipartial $r = -0.20$).[1] As can be seen in Fig. 1, simple slopes one standard deviation above and below the mean indicated that when free will beliefs were low, higher punitiveness predicted heightened anxiety ($b = 0.23$), $t = 5.89$, $p < 0.001$, but when free will beliefs were high, punitiveness had virtually no relationship with anxiety ($b = -0.003$), $t = -0.07$, $p = 0.95$.

2.3. Discussion

Free will beliefs moderated the impact of trait-level punitiveness on recent anxious feelings. Overall, there was a significant positive relationship such that more punitive people reported experiencing more symptoms of anxiety in the past week. This effect, weak across the full sample, was completely absent among people who believed rather strongly in free will, but quite pronounced among people who were most skeptical about free will. Thus, wanting to punish others while not believing that they freely choose to misbehave is associated with higher anxiety. Although these data are correlational, they do provide initial support for the view that believing in free will is an effective antidote to the distress associated with punishing people.

3. Study 2

At the individual difference level, Study 1 provided evidence linking punitive attitudes with anxiety as a function of free will beliefs. Study 2 (which, although in second position, was actually conducted after experimental evidence had been found) is another validation study, this time at the societal level, or more precisely with U.S. states as the unit of analysis. It investigated the relationships between free will belief, real world punitiveness (as measured by incarceration rates), and real world mental and emotional distress (as measured by rates of mental health issues) on a state-level. Such data inevitably lack the advantages of laboratory experimentation and invite alternative explanations, but they can certainly render a hypothesis implausible. If punishing contributes to psychological distress, then states with large prison populations might have high rates of mental health problems. If belief in free will mitigates that relationship, then the strength of that relationship would vary inversely with general belief in free will.

[1] Without covariates, punitiveness remained a significant predictor of higher anxiety, $\beta = 0.17$, $t = 3.25$, $p = 0.001$, 95% CI [0.038, 0.151], free will belief remained a significant predictor of lower anxiety, $\beta = -0.13$, $t = -2.51$, $p = 0.012$, 95% CI [−0.207, −0.025], and the interaction remained virtually unchanged, $\beta = -0.22$, $t = -4.63$, $p < 0.001$, 95% CI [−0.223, −0.092].

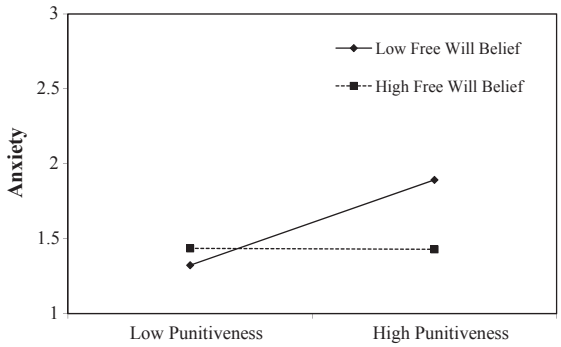

Fig. 1. Interaction between free will belief and punitiveness on the five-point anxiety symptom index, controlling for age, sex, education, religiosity, and ideology in Study 1.

One might argue that only a very small proportion of individuals are involved in the criminal justice system, so it may seem odd that the general population would suffer psychological consequences from the increased punitiveness of a very small group of individuals. However, a growing body of research shows that people often do not support the punishment of norm-violators (Eriksson, Strimling, & Ehn, 2013; Kiyonari & Barclay, 2008), that people negatively evaluate those who punish others (Eriksson, Andersson, & Strimling, 2015; Strimling & Eriksson, 2014), and that people will often punish those who punish others (Cinyabuguma et al., 2006). Consider, for example, longstanding negative attitudes toward police, or the pejorative terms "rat", "snitch", or "fink" for informers. That people tend not to support punishment suggests the likelihood that there would also be negative affective consequences to the awareness that one's own group members are being punished. Indeed, individuals living in zip codes with high rates of prison admissions experience higher rates of mental health issues (controlling for a variety of other risk factors), even among those who were never incarcerated themselves (Hatzenbuehler, Keyes, Hamilton, Uddin, & Galea, 2015). According to our hypothesis, this relationship ought to be even stronger when incarcerated others are perceived as not personally responsible for their crimes (i.e., when free will beliefs are low).

On a practical level, it is clear that individuals outside of the criminal justice system have interest in ensuring that only those who had the capacity to freely choose to perform a harmful behavior (and did so) are punished (e.g., Campaign for Youth Justice, The Innocence Project). Furthermore, previous work has found support for the relationship between higher exposure to criminal behavior on a country-level and higher country-level free will beliefs (Clark et al., 2014). Overall then, there is reason to believe that awareness of higher rates of incarceration would be related to higher free will beliefs, and that we would find the psychologically beneficial aspects of this relationship on this more macro-level. This of course assumes that people are reasonably aware of incarceration rates in their own communities. Almost certainly, people generally are not aware of factual numeric rates of incarceration. Still, higher incarceration rates may reverberate through the population insofar as more people will have at least indirect contact with the criminal justice system, such as by knowing someone involved (victims, perpetrators, jurors, judges, lawyers, police officers, prison employees, social workers) or by media exposure (to crimes, police action, court outcomes).

Of course, broad data of this sort are highly susceptible to confounding variables, which could potentially lead us to draw erroneous inferences based on the relationships between these variables. We sought to reduce this worry in three ways. First, we controlled for all of the same variables as in Study 1: age, sex, education, political ideology, and religiosity, as well as additional variables that could be related to free will belief, incarceration rates, and mental health issue rates. We controlled for crime rates because this would likely be positively related to both mental health issue rates and incarceration rates, and based on past research, negatively related to free will beliefs (e.g., Baumeister, Masicampo, & DeWall, 2009; Vohs & Schooler, 2008). Hence Study 2 is not simply a comparison of high-crime versus low-crime states. We also controlled for gross state product as this may be negatively related to incarceration rates and mental health issue rates. Last, we also controlled for corrections expenditures as this is likely highly related to incarceration rates, and based on past research, possibly also related to free will beliefs (e.g., Shariff et al., 2014).

Second, we constructed this set list of control variables prior to running any analyses. That is, we decided which variables should be included in our model as the best test of our hypothesis, as described above, and then ran our analyses just once controlling for those variables, without testing alternate models. This was done to reduce 'researcher degrees of freedom', eliminating our ability to support our hypothesis with *any* of the many conceivable models, rather than the model we decided a priori would be the best test of our hypothesis (Simmons, Nelson, & Simonsohn, 2011).

Third, we conducted the same analyses replacing incarceration rates as a predictor with similar, but less-punitive constructs: parole rates and probation rates. These rehabilitative alternatives to incarceration reflect specific outcomes for people who commit crimes, are caught for committing those crimes, and are held accountable for those crimes—but are much less severe in terms of general punitiveness. Incarcerated individuals may receive parole after being released from prison as a discontinuance of punishment, under the expectation that they will comply with a set of conditions. Despite that parolees were prisoners in the past (and are likely living in the same states, with the same characteristics [i.e., mental health issue rates, free will beliefs], in which they were incarcerated), we did not expect any relationship between parole rates and mental health issues, nor that state-level free will beliefs would moderate that relationship. Though similar to parole in treatment, probation is given as a contingent alternative to incarceration. Probation is essentially giving a person a second chance before making the decision to punish them (send them to

prison), and so, is even somewhat forgiving in nature. For this reason, we did not expect any relationship between probation rates and mental health issues, nor that state-level free will beliefs would moderate this relationship. In plain terms, we thought people would feel distress about sending large numbers of their fellow citizens to prison, but not about putting them on probation or parole.

3.1. Method

To obtain state-level free will beliefs, the responses of all participants (n = 4868; M_{age} = 37.81; 1684 female) from yourmorals.org, who completed the FAD (Paulhus & Margesson, 1994), were averaged by state (derived from IP addresses). The control variables age, sex, education, religiosity, and political ideology were measured the same way as in Study 1 and were also averaged by state.

Incarceration rates, parole rates, and probation rates by state (all per 100,000 in the population) were drawn from the United States Bureau of Justice Statistics. We also controlled for gross state product (GSP) drawn from the United States Department of Commerce Bureau of Economic Analysis, and percentage of state expenditures spent on corrections (corrections expenditures) drawn from the National Association of State Budget Officers. Finally, we controlled for crime rates by state (per 100,000 in the population), which were drawn from the Federal Bureau of Investigation. Both total violent crime rates and total property crime rates by state were converted to z-scores and then averaged. All predictor variables (except the interaction terms between free will beliefs and each of the following: incarceration rates, parole rates, and probation rates) were standardized, as the scales varied greatly in size (e.g., 0.2–0.83 for sex, 31,985–72,281 for GSP). Our dependent variable, rates of mental health issues by state (percentage of the population having any diagnosable mental, behavioral, or emotional disorder), was drawn from the Substance Abuse and Mental Health Services Administration. The majority of the yourmorals.org participants completed the FAD in the year 2012; for this reason, we utilized the available data from 2012 for all of these state-level variables (incarceration rates, parole rates, probation rates, gross state product, corrections expenditures, crime rates, and mental health issue rates).

3.2. Results

3.2.1. Correlations

Before we conducted our primary analyses, we analyzed the bivariate correlations between free will beliefs, prison rates, probation rates, parole rates, GSP, corrections expenditures, crime rates, and mental health issues. Here we will only report the significant correlations, but a full correlation matrix is available from the authors by request. Higher state-level free will beliefs were associated with higher incarceration rates, $r = 0.38$, $p = 0.009$. This is consistent with past work showing that both decreasing free will beliefs decreases punitiveness (Shariff et al., 2014) and that increasing punitiveness increases free will beliefs (Clark et al., 2014). This demonstrates that our state-level estimates of free will belief based on *yourmorals.org* participants map onto our state-level punitiveness variable based on entire state populations in the way that would be predicted by past research addressing similar issues to the present research. Not surprisingly, higher rates of crime were positively associated with higher rates of incarceration, $r = 0.54$, $p < 0.001$, simply indicating that states with more crime incarcerate more people. Only higher GSP was related to lower rates of mental health issues, $r = -0.52$, $p < 0.001$. Higher GSP was also related to lower rates of incarceration, $r = -0.36$, $p = 0.012$. These two results simply indicate that more financially successful states are also mentally healthier and have fewer prisoners. The only remaining significant relationship was that higher rates of probation were associated with lower corrections expenditures, $r = -0.34$, $p = 0.017$, perhaps because probation is an economical alternative to incarceration.

3.2.2. Incarceration

The main analysis examined links among free will belief, incarceration, and mental distress. Mental health issue rates were regressed on free will beliefs, incarceration rates, and the interaction controlling for age, sex, education, religiosity, political ideology, GSP, corrections expenditures, and crime rates. Three states (Illinois, Nevada, and Washington) could not be included in the analysis because incarceration rates were not reported for those three states in 2012. As can be seen in Table 2, state-level free will beliefs had no significant effect on rates of mental health issues, $\beta = -0.02$, $t = -0.08$, $p = 0.938$, 95% CI [−0.280, 0.259], partial $r = -0.01$, semipartial $r = -0.01$. Higher incarceration rates (non-significantly) predicted higher rates of mental health issues, $\beta = 0.34$, $t = 1.63$, $p = 0.113$, 95% CI [−0.047, 0.421], with a small to medium effect size, partial $r = 0.27$, semipartial $r = 0.19$.

Consistent with the main hypothesis, there was a significant interaction, $\beta = -0.36$, $t = -2.34$, $p = 0.025$, 95% CI [−0.296, −0.021] between incarceration rates and free will beliefs on rates of mental health issues, with a medium effect size, partial $r = -0.37$, semipartial $r = -0.28$.[2,3] As depicted in Fig. 2, simple slopes one standard deviation above and below the mean indicated that in states where free will beliefs were low, higher incarceration rates predicted higher rates of mental health issues ($b = 0.35$), $t = 2.15$, $p = 0.039$, but in states where free will beliefs were high, incarceration rates were unrelated to rates of mental health issues ($b = 0.03$), $t = 0.27$, $p = 0.789$.

[2] Incarceration rates include imprisonment of US residents of all ages. If only 18 and older are included, the interaction between incarceration rates and free will beliefs on rates of mental health issues remains significant, $\beta = -0.36$, $t = -2.32$, $p = 0.026$.

[3] Without covariates, state-level free will beliefs had no effect on mental health issue rates, $\beta = -0.08$, $t = -0.48$, $p = 0.635$, 95% CI [−0.217, 0.134], semipartial $r = -0.07$. Higher incarceration rates significantly predicted higher rates of mental health issues, $\beta = 0.17$, $t = 3.25$, $p = 0.001$, 95% CI [0.038, 0.151], with a small to medium effect size, semipartial $r = 0.23$. The interaction fell out of statistical significance, $\beta = -0.25$, $t = -1.46$, $p = 0.152$, 95% CI [−0.266, 0.043], but maintained a small to medium effect size, semipartial $r = -0.22$.

Table 2
Mental health issues rates regressed on incarceration rates, parole rates, or probation rates, state-level free will belief, the associated interaction, and relevant controls in Study 2.

	F	R^2	*b*	*SE*	*t*	*p*
Mental Health Rates (Model)	3.29	0.51				0.003
Age			−0.07	0.09	−0.80	0.428
Sex			0.15	0.09	1.78	0.083
Education			−0.08	0.09	−0.84	0.406
Religiosity			−0.05	0.09	−0.53	0.603
Political Ideology			−0.09	0.13	−0.66	0.516
GSP			−0.23	0.08	−3.09	0.004
Corrections Expenditures			−0.15	0.08	−1.84	0.074
Crime Rates			0.03	0.10	0.27	0.786
Incarceration Rates			0.19	0.12	1.63	0.113
Free Will Belief			−0.01	0.13	−0.08	0.938
Incarceration × Free Will Belief			**−0.16**	**0.07**	**−2.34**	**0.025**
Mental Health Rates (Model)	2.15	0.38				0.040
Age			−0.04	0.11	−0.36	0.724
Sex			0.10	0.08	1.15	258
Education			−0.10	0.10	−1.01	0.321
Religiosity			−0.03	0.09	−0.32	0.752
Political Ideology			−0.05	0.15	−0.36	0.721
GSP			−0.28	0.08	−3.46	0.001
Corrections Expenditures			0.00	0.08	−1.09	0.285
Crime Rates			0.06	0.10	0.62	0.538
Parole Rates			−0.06	0.08	−0.82	0.416
Free Will Belief			−0.07	0.14	−0.46	0.650
Parole × Free Will Belief			**−0.08**	**0.09**	**−0.84**	**0.409**
Mental Health Rates (Model)	2.67	0.44				0.012
Age			0.02	0.11	0.21	0.834
Sex			0.10	0.08	1.19	0.243
Education			−0.08	0.10	−0.83	0.411
Religiosity			0.01	0.09	0.12	0.904
Political Ideology			−0.09	0.14	−0.60	0.549
GSP			−0.25	0.08	−3.33	0.002
Corrections Expenditures			−0.13	0.09	−1.48	0.148
Crime Rates			0.06	0.10	0.64	0.529
Probation Rates			−0.05	0.10	−0.57	0.576
Free Will Belief			−0.02	0.14	−0.13	0.900
Probation × Free Will Belief			**0.17**	**0.10**	**1.71**	**0.095**

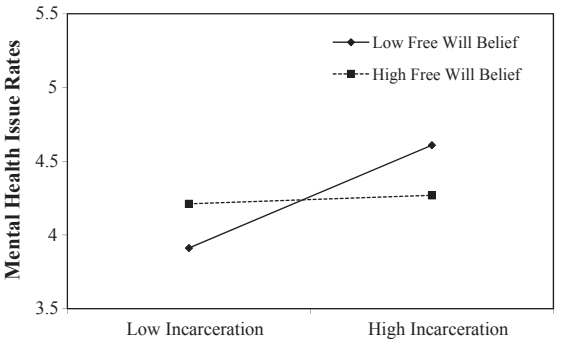

Fig. 2. Interaction between state-level free will belief and incarceration rates on mental health issue rates, controlling for relevant variables in Study 2.

3.2.3. Parole

A second analysis replaced incarceration with parole. There was no relationship between state-level free will beliefs and mental health issues, $\beta = -0.12$, $t = -0.46$, $p = 0.650$, 95% CI [−0.354, 0.223], partial $r = -0.07$, semipartial $r = -0.06$ (see Table 2). Parole rates did not significantly predict mental health issue rates, $\beta = -0.11$, $t = -0.82$, $p = 0.42$, 95% CI [−0.222, 0.094], partial $r = -0.13$, semipartial $r = -0.11$, nor did the interaction, $\beta = -0.12$, $t = -0.84$, $p = 0.409$, 95% CI [−0.260, 0.108], partial $r = -0.13$, semipartial $r = -0.11$.[4]

[4] Without covariates, both main effects and the interaction remained non-significant, *ps* > 0.627.

3.2.4. Probation

Last, we repeated the same analyses using probation rates instead of parole or incarceration. There was no relationship between state-level free will beliefs and mental health issues, $\beta = -0.03$, $t = -0.13$, $p = 0.900$, 95% CI [−0.296, 0.261], partial $r = -0.02$, semipartial $r = -0.02$; and no relationship between probation rates and mental health issues, $\beta = -0.09$, $t = -0.57$, $p = 0.58$, 95% CI [−0.245, 0.138], partial $r = -0.09$, semipartial $r = -0.07$ (Table 2). However, unexpectedly, there was a marginal interaction between probation rates and free will beliefs, $\beta = 0.25$, $t = 1.71$, $p = 0.095$, 95% CI [−0.032, 0.380], with a small to medium effect size, partial $r = 0.27$, semipartial $r = 0.21$, demonstrating the opposite pattern of the interaction between incarceration rates and free will beliefs.[5] As can be seen in Fig. 3, simple slopes one standard deviation above and below the mean indicated that when state-level free will beliefs were high, there was no relationship between probation rates and mental health issues ($b = 0.12$), $t = 0.74$, $p = 0.466$, but when free will beliefs were low, higher rates of probation were associated with fewer mental health issues ($b = -0.23$), $t = -2.16$, $p = 0.037$. Put another way, states with high rates of this more lenient sentence actually seemed to have *reduced* rates of mental health issues when the citizens of those states had relatively low levels of belief in free will.

3.3. Discussion

Utilizing ecologically valid measures, Study 2 demonstrated results consistent with the main hypothesis that free will beliefs help alleviate the mental and emotional upset that can result from the punishment of fellow humans. Specifically, Study 2 demonstrated that high state-level rates of incarceration predicted higher rates of mental health issues—but only for states that held, on average, relatively low levels of belief in free will. The relationship between incarceration and mental health issues dropped to almost zero in states with relatively high belief in free will.

These results should be interpreted with caution, as many factors can influence these broad-level constructs. Still, these results are not due to differences in crime rates, poverty, education, religiosity, or state prison budgets. Though we controlled for the obvious relevant variables, there is always the possibility that one or many unknown factors are accounting for the effect. Furthermore, ideally, our measure of free will belief would have been based on entire state populations as were our measures of mental health and punitiveness. To our knowledge, free will belief data are not available for entire state populations, and so, we resorted to using state-level estimates of free will beliefs based on *yourmorals.org* respondents. However, the predicted relationship emerged even with this relatively small sample size (47–50 states), with rough estimates of state-level free will beliefs, controlling for all potential confounds we could generate. Confidence in these results is bolstered by the finding that our measure of state-level free will beliefs was positively associated with incarceration rates, consistent with previous evidence linking free will beliefs to punishing (Clark et al., 2014; Shariff et al., 2014).

Moreover, the significant interaction between free will beliefs and incarceration on mental health was specific to harsh punishment: We did *not* find the same interaction between free will beliefs and either probation or parole on mental health outcomes. (In fact, probation showed a marginal interaction in the opposite direction.) Thus, the community's overall distress seems linked to how frequently it uses severe punishment (imprisonment), rather than simply reflecting how many citizens are involved in the criminal justice system in less severe ways. These results may suggest that it is really this more tangible, harsh form of punishment that leads people to experience mental and emotional distress when criminal behavior is viewed as relatively externally determined.

Although we acknowledge the limitations of this study due to the highly complex nature of our constructs of interest, and caution against overreliance on this specific set of results, the consistency of these results with Study 1 (and the upcoming Studies 3–5) suggests a realistic possibility that our interpretation is correct: believing in free will may have positive benefits for emotional and mental well-being in coping with the realization that one either passively endorses or actively promotes the locking up of their fellow citizens.

4. Study 3

We now shift to report laboratory studies, which can test causal relationships and reduce potential confounds that inevitably attend the correlational designs of Studies 1–2. Study 3 used an experimental and quasi-experimental design to test the hypothesis that punishment can be distressing, particularly in situations where the punished other had limited control. We tested this in a realistic context by having participants receive an unfair offer from an ostensibly real partner in an economic game and decide whether to punish their partner. We then manipulated how free their partner was to make the unfair offer and measured punitive distress on both a general anxiety measure and a measure specific to their feelings about their choice to punish. We predicted that participants who chose to punish would experience more distress than those who refrained from punishing—but only when their partner did not freely choose to treat them unfairly.

[5] Without covariates, state-level free will beliefs had no effect on mental health issue rates, $\beta = -0.06$, $t = -0.43$, $p = 0.672$, 95% CI [−0.195, 0.127], semipartial $r = -0.06$. Higher probation rates were unrelated to mental health issues, $\beta = 0.04$, $t = 0.29$, $p = 0.777$, 95% CI [−0.153, 0.203], semipartial $r = 0.04$. However, the interaction became significant, $\beta = 0.34$, $t = 2.19$, $p = 0.034$, 95% CI [0.019, 0.453], maintaining a medium effect size, semipartial $r = 0.31$.

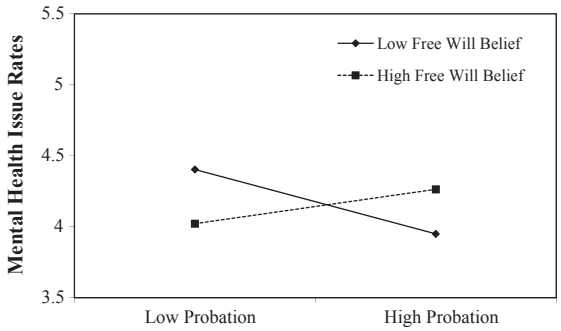

Fig. 3. Interaction between state-level free will belief and probation rates on mental health issue rates, controlling for relevant variables in Study 2.

4.1. Method

Study 3 was preregistered for sample size, analyses, and predictions (https://aspredicted.org/wxckq.pdf). Two hundred and eighty-eight Amazon's mechanical turk workers (M_{age} = 32.67; 161 female) participated in an online experiment in exchange for a small payment. From a pre-test with twenty-one participants, we anticipated that approximately 45% of participants would make the decision to punish; in aiming for at least 60 participants per condition, we posted 280 timeslots. Due to multiple people signing up at the same time, 288 participants managed to sign up. In a 2 (punish vs. no punish) × 2 (choice vs. no choice) between-subjects design, all participants were informed that they would be participating in two decision-making tasks partnered with two different workers (one partner for each task). In each task, there would be $2.00 to split between the partners. In the first task, their partner would select how to allocate the money between the two of them; in the second task, they would select how to allocate the money. All participants then proceeded to Round 1, in which their partner "jake2880" would choose how to allocate the money. After a short delay, participants were informed that jake2880 allocated $0.30 to them and $1.70 to himself.

Participants were informed that the average partner allocation up to that point had been $0.87, and therefore they had received $0.57 less than average (thus barely a third of the typical allocation). For purposes of appearing as though we were interested in their reaction to their offer, they were then asked to respond to a variety of questions regarding the fairness and generosity of their partner's allocation and the extent to which they were pleased or angry. These questions were irrelevant to our predictions, and so were not analyzed (as specified in the preregistration). Participants were then told:

> "For all workers who were allocated less than $0.50, we are providing the opportunity to even the payout between the offerer and the receiver. Because you were allocated only **$0.30,** you can reduce the payment of **jake2880** to match your own payment. In your case, you can reduce **jake2880**'s payment from **$1.70** to **$0.30** so you will both earn **$0.30**. **jake2880** will not be informed of your decision until the conclusion of the study. Would you like to lower **jake2880**'s payment to **$0.30**?"

Participants then selected whether to punish jake2880 by reducing his payment to match their own. Following this decision, participants were told that because jake2880 was the offerer in this round, he had the opportunity to share a message with them. After a delay, participants were randomly assigned to receive one of two messages:

> **No choice condition:** "i think something went wrong with this survey bcuz it told me in this round that i would get to be an "offerer" and choose an amount to give to the "receiver" but the scrolling bar was stuck at 30 cents. i had no choice but to offer 30 cents"
> **Choice condition:** "i think something went wrong with this survey bcuz i had to click to offer 30 cents like three times before it would let me click next. i eventually got it to work though"

This choice manipulation purposefully came after participants made the decision to punish so as not to influence their decision about whether to punish. As a measure of their general anxiety, participants then completed a shortened version of the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970) adapted by Marteau and Bekker (1992). This scale consists of 6 items measuring state anxiety levels (e.g., "I feel calm.", "I feel tense."), each rated on a 7-point scale from "Not at all" to "Very much so" (α = 0.90). To furnish a measure of distress specific to the act of punishment, participants were reminded of their decision to either reduce jake2880's payment or not, and then reported how they felt about their choice on six items (satisfied, guilty, pleased, positive, anxious, bad) on 7-point scales from "Not at all" to "Very much", scored such that higher values indicated greater punitive distress (α = 0.90). When it came time to complete Round 2, participants were told that all other workers had already been assigned partners, and that they were excused from completing Round 2.

Multiple steps were taken to increase the believability of the decision-making task. First, participants were told that they must complete the study in one sitting so as not to delay their partner's participation. Participants also selected their own username for participating in the decision-making task, to which they were referred throughout the rest of the study (so as to be comparable to "jake2880"). There was a delay before participants were assigned their partner and each time their partner was making a decision or typing a response. All pronouns referring to jake2880 were gender neutral so as to indicate we were not previously aware of who their

partner would be. Participation was also limited to workers who had completed fewer than 500 HITs to minimize the likelihood that they would have been deceived by similar economic games in the past. Nonetheless, 40 participants expressed some degree of doubt about whether jake2880 was a real person in a suspicion probe at the end of the study. Those who chose to punish were no less suspicious (11.5%) than those who chose not to punish (15.8%), $\chi^2 = 1.10$, $p = 0.295$. Those who believed jake2880 had no choice (14.0%) were equally as suspicious as those who believed jake2880 freely chose the unfair offer (13.8%), $\chi^2 = 0.002$, $p = 0.962$. Excluding suspicious participants did not alter the statistical significance of any main effects, interactions, or simple effects discussed below.

4.2. Results

4.2.1. Punishment

As expected from the pre-test, approximately 45% of people chose to punish ($n = 130$).

4.2.2. Distress

A 2 × 2 multivariate analysis of variance (MANOVA) on the two distress measures revealed no main effect for choice condition on general anxiety, $F(1,284) = 0.74$, $p = 0.392$, $\eta^2_p = 0.003$, nor punitive-specific distress, $F(1,284) = 1.52$, $p = 0.278$, $\eta^2_p = 0.004$. There was a small to medium main effect of punishment on general anxiety such that participants who chose to punish were more anxious ($M = 2.88$, $SD = 1.36$) than those who chose not to punish ($M = 2.40$, $SD = 1.19$), $F(1,284) = 10.68$, $p = 0.001$, $\eta^2_p = 0.036$; and a similar large main effect of punishment on punitive-specific distress such that participants who chose to punish experienced far more distress over their decision ($M = 3.26$, $SD = 1.33$) than those who chose not to punish ($M = 1.97$, $SD = 0.99$), $F(1,284) = 93.48$, $p < 0.001$, $\eta^2_p = 0.248$. These results support the (occasionally contested) notion that punishment can be distressing.

Most important, there were significant choice × punishment interactions on both general anxiety, $F(1,284) = 6.05$, $p = 0.015$, $\eta^2_p = 0.021$, and punitive-specific distress, $F(1,284) = 15.69$, $p < 0.001$, $\eta^2_p = 0.052$, with small to medium effect sizes.[6] For participants who believed their partner freely chose the unfair offer, there was no difference in general anxiety between those who punished ($M = 2.76$, $SD = 1.23$) and those who did not ($M = 2.64$, $SD = 1.23$), $F(1,284) = 0.33$, $p = 0.568$, $\eta^2_p = 0.001$. However, for participants who believed their partner did not freely choose the unfair offer, participants who chose to punish ($M = 3.00$, $SD = 1.47$) were significantly more anxious than those who chose not to ($M = 2.14$, $SD = 1.09$), $F(1,284) = 16.39$, $p < 0.001$, $\eta^2_p = 0.055$. For punitive-specific distress, among participants who believed their partner freely chose the unfair offer, those who chose to punish were significantly more distressed with their decision ($M = 2.92$, $SD = 1.15$) than those who did not ($M = 2.15$, $SD = 1.12$), $F(1,284) = 16.29$, $p < 0.001$, $\eta^2_p = 0.054$, suggesting that punishing others can be distressing even when the punished other freely chose their actions. However, this difference was much larger among those who believed their partner did not freely choose the unfair offer ($M = 3.59$, $SD = 1.42$ for punishers; $M = 1.77$, $SD = 0.79$ for non-punishers), $F(1,284) = 92.86$, $p < 0.001$, $\eta^2_p = 0.246$.

4.3. Discussion

Participants who chose to punish an unfair other experienced greater general anxiety and punitive-specific distress than participants who chose not to punish, but only when they came to learn that their partner did not freely choose the unfair offer. When participants believed their partner freely chose to allocate them an unfair sum of money, their decision to punish had little if any effect on their distress levels. In a fairly realistic context, where people believed they were actually punishing another person, these results demonstrated that punishment is distressing in situations where the punished other had limited control over their harmful behavior.

Study 3 manipulated the partner's ostensible freedom of choice. That is, participants were randomly assigned to learn either that the partner had deliberately chosen the unfair allocation or had been forced to by a computer glitch. The decision whether to punish, however, was not randomly assigned, and instead participants self-selected into the punish and no punish conditions. Those who chose to punish may have differed in meaningful ways (such as general punitiveness) from those who chose not to punish. Still, punitive individuals being more anxious in general might explain the main effect of punishment on distress, but it is not clear how this difference could explain the significant interaction between punishment and free choice.

5. Study 4

Study 4 tested our hypothesis experimentally by manipulating belief in free will and then measuring punitive distress. First, punitive motives were activated by having participants read about a harmful behavior (Clark et al., 2014). Participants were then randomly exposed to an argument opposing or supporting the existence of free will, or free will beliefs were left at their (presumably high) baseline. We predicted that the anti-free will argument would increase punitive distress relative to the no argument and the pro-

[6] In the 21 participant pre-test, these same patterns emerged with medium/medium-large effects for both general anxiety, $F(1,20) = 2.26$, $p = 0.151$, $\eta^2_p = 0.118$, and punitive-specific distress, $F(1,20) = 1.99$, $p = 0.176$, $\eta^2_p = 0.105$ (though of course not significantly so due to the tiny sample size), further boosting our confidence in the replicability of these findings.

argument conditions. In contrast, because people normally increase their free will beliefs after exposure to an immoral action (Clark et al., 2014), participants who read an argument that supports free will (thus justifying their desire to punish) would presumably not experience heightened or relieved distress relative to the control group.

Study 4 also included measures of perceptions of the specific perpetrator's free will as well as more general belief in free will to determine whether reduced perceived free will mediates the influence of the anti-free will condition on punitive distress.

5.1. Method

One hundred and sixty-nine undergraduates (M_{age} = 20.49; 134 female) participated in an online experiment in exchange for course credit. We aimed for approximately 60 participants per condition (180 total); 11 participants failed to complete the study procedures, resulting in 169. To induce punitive motives, all participants first read a scenario about a home robbery (adapted from Clark et al., 2014):

> *Sam, a special education teacher, wakes up one morning and finds that someone robbed and vandalized his home while he was sleeping. His window is broken, his house is trashed, and all of his valuables are missing.*

Participants were then randomly assigned to read a pro-free will argument, an anti-free will argument, or no argument. Participants in the argument conditions were told they would read a paragraph describing the latest scientific opinion on free will (written by the lead author; 166 – 172 words long; full arguments available in supplementary materials). The anti-free will argument opposed the existence of free will (e.g., "*Recent research in psychology and cognitive neuroscience suggests that free will is an illusion. Brain scanning technology has allowed scientists to observe that neural evidence of behavior actually precedes conscious awareness of the behavior… science is coming closer and closer to accounting for all of the variance in human behavior…*"). The pro-free will argument asserted that free will does exist (e.g., "*Recent research in psychology and cognitive neuroscience suggests that humans actually do have free will. Brain scanning technology has allowed scientists to observe that the conscious experience of deciding to perform a behavior does indeed precede the behavior itself… scientists are realizing that it will never be possible to predict human behavior…*").

Participants then rated the robber's free will by responding to three items (whether the behavior was freely chosen, whether the robber could have made other choices, and whether the robber exercised their own free will) on 7-point scales from "Not at all" to "Very much so" (α = 0.75). We expected that participants who read the anti-free will argument would attribute less free will to the robber than participants in the no argument and pro-free will argument conditions. However, we did not expect participants who read the pro-free will argument to attribute more free will to the robber than those in the no argument condition, because baseline free will beliefs tend to be high, and people naturally increase their free will attributions and beliefs following exposure to immoral behavior (i.e., there would be a ceiling effect). It is typically the case in free will research that pro-free will arguments and neutral controls produce the same result, while anti-free will arguments will produce different results from both (Vohs & Baumeister, 2009).

As our main dependent measure, participants then reported their punitive distress on two items: how they would feel about the robber spending time in prison and how they would feel if they were the judge who sentenced the robber to prison on 7-point scales from "Very bad" to "Very good" (r = 0.81, $p < 0.001$). Finally, participants reported their free will beliefs on the Free will and Determinism-Plus Scale (FAD-Plus; Paulhus & Carey, 2011; adapted from the FAD scale used in Studies 1–2), which contains a subscale measuring free will belief on seven items (e.g., "People have complete free will."), each rated on a 5-point scale from "Strongly disagree" to "Strongly agree". We predicted that participants who read the anti-free will argument would experience heightened punitive distress relative to participants in the no argument and pro-free will argument conditions, due to their reduced perceptions of free will. Because experiencing punitive motives leads people to naturally increase their belief in free will, this should naturally dampen the associated distress, thus we expected little or no difference in punitive distress between participants in the no argument and pro-free will argument conditions.

5.2. Results

5.2.1. Manipulation check

The anti-free will argument had a small to medium (though non-significant) effect on free will attributions, $F(2,162)$ = 2.23, p = 0.110, η^2_p = 0.027. Participants who read the anti-free will argument attributed significantly less free will to the perpetrator (M = 5.34, SD = 1.20) than participants who read no argument (M = 5.78, SD = 1.09), p = 0.047, but not quite significantly less than those who read the pro-free will argument (M = 5.65, SD = 1.07), p = 0.138. As expected, participants who read the pro-free will argument did not differ from participants who read no argument, p = 0.576.

The anti-free will argument also had a small to medium (non-significant) effect on general free will beliefs, $F(2,163)$ = 2.00, p = 0.139, η^2_p = 0.024. Participants who read the anti-free will argument believed marginally less in free will (M = 3.49, SD = 0.56) than participants who read no argument (M = 3.70, SD = 0.56), p = 0.063, though once again not quite significantly less than those who read the pro-free will argument (M = 3.65, SD = 0.61), p = 0.142. As expected, participants who read the pro-free will argument did not differ from participants who read no argument, p = 0.652.

5.2.2. Punitive distress

An Analysis of Variance (ANOVA) revealed that the argument condition had a small to medium significant effect on punitive distress, $F(2,166)$ = 3.38, p = 0.037, η^2_p = 0.039. Participants who read the anti-free will argument (M = 3.88, SD = 1.10) felt significantly worse than participants in the pro-free will condition (M = 4.47, SD = 1.18), p = 0.012, but not quite significantly

worse than participants who read no argument ($M = 4.27$, $SD = 1.58$), $p = 0.113$. As anticipated, participants who read the pro-free will argument felt no better or worse than participants who read no argument, $p = 0.416$.

5.2.3. Mediation

To determine whether reduced belief in free will mediated the influence the condition on punitive distress, we conducted a bootstrap mediation analysis (5000 resamples) using the MEDIATE macro (Hayes & Preacher, 2014), with the no argument condition specified as the reference category.[7] As predicted, reduced belief in free will in the anti-free will argument condition (relative to the no argument condition) accounted for participants' heightened punitive distress, bias-corrected bootstrap 95% CI [−0.352, −0.006]. As expected, the indirect effect of the pro-free will argument on punitive distress through free will belief did not emerge, bias-corrected bootstrap 95% CI [−0.217, 0.084].

When free will attributions were replaced as the mediator, we similarly found that reduced perceptions of the specific perpetrator's free will accounted for higher punitive distress in the anti-free will condition (relative to the no argument condition), bias-corrected bootstrap 95% CI [−0.335, −0.013]. Once again, the indirect effect of the pro-free will argument on punitive distress through attributions of free will did not emerge, bias-corrected bootstrap 95% CI [−0.196, 0.081].

5.3. Discussion

Study 4 found that participants reported more distress over punishing a hypothetical criminal if their beliefs in free will had been weakened, as opposed to bolstered. That disbelief in free will increases punitive distress supports our contention that believing in free will helps alleviate distress over punishing. Study 4 used a purely experimental design and hence does not suffer from the ambiguities associated with correlational and self-selection designs, as in Studies 1–3.

Two limitations must be acknowledged. First, the control condition (no free will argument) fell in between the other two, such that it did not differ significantly from either. Second, the manipulation checks registered only weak effects of the manipulations on free will beliefs and attributions. That the manipulations affected the dependent measures more strongly than the manipulation checks could indicate that the manipulation checks were rather insensitive. Still, given these weaknesses, Study 5 was designed as an improved replication.

6. Study 5

Study 5 improved upon Study 4 in a number of ways. One design change was to add a condition in which no crime or misdeed was committed, so there would be no impulse to punish. This enabled us to assess any possible direct impact of the free will belief manipulation on distress. Study 4 found more anxiety in the anti-free will condition than the pro-free will condition, which we interpreted in connection with punishment, but in principle it could have stemmed directly from having one's belief in free will undermined.

A second question was whether the combination of punitive desires and reduced belief in free will would indeed produce a general sort of distress, as opposed to merely activating a reluctance to punish. Past work has shown that reducing free will beliefs reduces punitive impulses (Shariff et al., 2014), so perhaps this accounted for Study 4's findings: It is possible that participants in Study 4 merely reported feeling bad about punishment as a way of expressing their disinclination to punish. That is, they might not have actually felt bad but merely used the item to make a point. Study 5 added a measure of general anxiety, which would not have that ambiguity.

Study 5 used two different approaches to test our hypothesis. First, we manipulated the desire to punish by having participants read about either a criminal or neutral behavior. Participants then read one of the same anti-free will or pro-free will arguments from Study 4. We predicted that the anti-free will argument would lead to heightened anxiety in the criminal condition compared to the neutral condition. In contrast, participants who read an argument supporting free will, thus justifying their desire to punish, should not experience higher anxiety in the criminal condition than the neutral condition. These results would support the hypothesis that punitive distress arises from the combination of punitive desires and disbelief in free will.

Our second approach involved the inclusion of a third argument condition in which it was said that punishment is justified as a means of deterring harmful behavior, regardless of whether free will does or does not exist. If people are motivated to see others as free in order to justify punishment, this argument should reduce attributions of free will, despite not indicating anything about whether free will actually exists. In other words, if people increase their free will beliefs after exposure to harmful actions as a means of justifying punitive impulses, providing this alternate punishment justification should reduce the necessity to perceive immoral actors as free. Further, because this condition justifies punitive impulses even in the absence of free will, it should have no effect on punitive impulses, but it should reduce the disparity in anxiety levels between participants who read about criminal and neutral behaviors similar to the pro-free will argument condition. Thus, for participants who read the "punish for deterrence" argument, we predicted that there would be no differences in attributions of free will and no differences in anxiety levels between participants experiencing heightened punitive motives and those not.

[7] Bootstrap mediation procedures are now a common practice for testing mediation as this approach does not assume normal distributions and therefore has increased power (e.g., Preacher & Hayes, 2004; Shrout & Bolger, 2002).

6.1. Method

Three hundred and ninety-four undergraduates (M_{age} = 20.72; 278 female) participated in an online experiment in exchange for course credit. We once again aimed for approximately 60 participants per condition (360 total), but knowing that some participants fail to complete online experiments after signing up, we posted 400 timeslots. Of the 400 participants who signed up, 394 completed the experimental procedures. In a 2 (behavior: criminal vs. neutral) × 3 (argument: pro-free will, anti-free will, punish for deterrence) between-subjects design, participants first read one of two hypothetical scenarios. In the criminal condition, participants read about the same harmful behavior from Study 4 (a man robbing a home); in the neutral condition, participants read about a neutral behavior (a man taking aluminum cans out of a recycling bin):

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.

To ensure that the criminal condition increased the desire to punish and attributions of free will to the perpetrator, after reading the scenario, participants rated the actor's free will by responding to the same three items from Study 4 (whether the behavior was freely chosen, whether the actor could have made other choices, and whether the actor exercised their own free will) on 7-point scales from "Not at all" to "Very much so" (α = 0.71), and reported the extent to which the actor should be punished on a 7-point scale from "Not at all" to "Very much so". These also served as pre-argument measures of free will attributions and punitiveness to ensure that the scientific arguments had their intended influence.

Participants were then told they would read a paragraph describing the latest scientific opinion on free will. They were randomly assigned to read one of three arguments. The pro-free will and anti-free will arguments were the same as in Study 4. The third argument condition (172 words, written by the lead author), the "punish for deterrence" argument, opposed the necessity of free will for punishment (e.g., "… *philosophers and scientists are realizing that free will is not a requirement for punishment… If humans know that bad behavior will be punished, they will be less likely to perform that bad behavior. Punishment is not an issue of free will or moral responsibility; it is merely a means of discouraging behavior destructive to society."*). By reducing the apparent necessity of free will for justifying punishment, we predicted that this condition would reduce attributions of free will, despite not indicating anything about whether free will actually exists. This result would suggest that one of the reasons people are compelled to see actions as free is a desire to justify their punitive impulses. Because this argument justifies punitive impulses even in the absence of free will, we also predicted that this condition would reduce the disparity in anxiety levels between those experiencing punitive motives and those who are not.

Immediately following the argument, participants reported their general anxiety on the STAI short form (as in Study 3; Marteau and Bekker,1992; α = 0.86). This was the primary dependent measure. Finally, as post-argument measures, participants were asked in light of the scientific information they read, how they rated the actor's free will from the earlier scenario (α = 0.84) and the extent to which the actor should be punished (on the same items as before the argument).

6.2. Results

6.2.1. Punitive motives

The behavior condition successfully manipulated punitiveness. Participants in the criminal condition had greater desire to punish (M = 6.22, SD = 1.24) than participants in the neutral condition (M = 2.97, SD = 1.61), $t(389) = -22.45$, $p < 0.001$, 95% CI [−3.526, −2.959].

We also analyzed the influence of the arguments on the desire to punish from before to after exposure to the free will arguments. A 3 × 2 mixed ANOVA with argument condition as the between subjects variable and pre-argument versus post-argument as the within subjects variable showed that there were differences among argument conditions in how much punitive motives changed from before to after the argument, $F(2, 388) = 8.43$, $p < 0.001$, with a small to medium effect size, $\eta^2_p = 0.042$. Consistent with past work, the desire to punish marginally increased after reading the Pro-Free Will Argument (Time 1 M = 4.55, SD = 2.13; Time 2 M = 4.71, SD = 2.11), $p = 0.060$, and significantly decreased after reading the Anti-Free Will Argument (Time 1 M = 4.88, SD = 2.17; Time 2 M = 4.56, SD = 2.13), $p < 0.001$. Though the "Punish for Deterrence" argument reduced attributions of free will (discussed below), this condition did not reduce participants' overall desire to punish (Time 1 M = 4.51, SD = 2.17; Time 2 M = 4.36, SD = 2.12), $p = 0.116$, suggesting that this argument may have succeeded in disentangling free will beliefs and the capacity for punishment. The behavior condition (criminal vs. neutral) did not moderate these effects, $ps > 0.18$.

6.2.2. Free will attributions

The argument manipulation also successfully influenced attributions of free will. Participants who read the argument opposing the existence of free will gave the lowest post-argument attributions of free will (M = 5.02, SD = 1.42), followed by participants who read the passage indicating that punishment is justified for deterrent purposes (M = 5.59, SD = 1.30), followed by participants who read the passage supporting the existence of free will (M = 6.10, SD = 1.12), $F(1, 388) = 24.55$, $p < 0.001$, $\eta^2_p = 0.112$. All conditions significantly differed from each other, $ps < 0.004$. We also analyzed whether free will attributions changed from before to after each of the arguments. A 3 × 2 mixed ANOVA with the argument condition as the between subjects variable and pre-argument and post-argument as the within subjects variable showed differences between argument conditions in how much free will attributions changed from before to after the argument, $F(2, 383) = 22.74$, $p < 0.001$, $\eta^2_p = 0.106$. As intended, free will attributions increased significantly after reading the Pro-Free Will Argument (Time 1 M = 5.85, SD = 1.10; Time 2 M = 6.10,

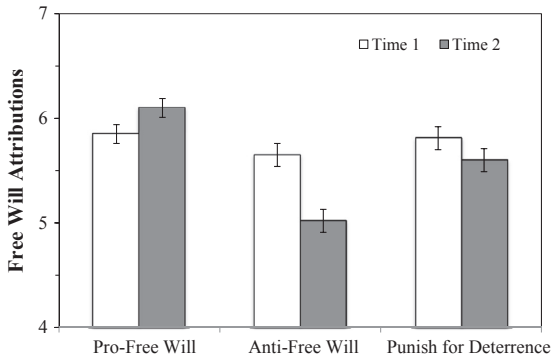

Fig. 4. Change in free will attributions from before to after each argument in Study 5.

$SD = 1.12$), $p = 0.009$, and decreased significantly after reading the Anti-Free Will Argument (Time 1 $M = 5.65$, $SD = 1.30$; Time 2 $M = 5.02$, $SD = 1.43$), $p < 0.001$.

Supporting our hypothesis, the "Punish for Deterrence" argument also decreased attributions of free will (Time 1 $M = 5.81$, $SD = 1.13$; Time 2 $M = 5.60$, $SD = 1.31$), $p = 0.048$, despite not indicating anything about whether free will does or does not exist (see Fig. 4). These results suggest that reducing the apparent necessity of free will for punishment reduces the perception that people act freely.

Consistent with past work, participants also attributed marginally more free will to the criminal behavior ($M = 5.88$, $SD = 1.13$) than the neutral behavior ($M = 5.65$, $SD = 1.24$), $t(386) = -1.91$, $p = 0.057$, 95% CI [−0.468, 0.007]. The behavior condition did not moderate any of these effects, $ps > 0.49$.

6.2.3. Anxiety

The main dependent measure was the anxiety report. An ANOVA revealed a significant argument × behavior interaction, $F(2,376) = 3.61$, $p = 0.028$, $\eta^2_p = 0.019$. As can be seen in Fig. 5, participants who read the Anti-Free Will argument were significantly more anxious in the criminal condition ($M = 3.16$, $SD = 1.12$) than the neutral condition ($M = 2.51$, $SD = 0.93$), $F(1,376) = 10.78$, $p = 0.001$, $\eta^2_p = 0.028$. There were no differences in anxiety levels between the criminal and neutral conditions in the Pro-Free Will, $F(1,376) = 0.52$, $p = 0.470$, $\eta^2_p = 0.001$, and "Punish for Deterrence" argument conditions, $F(1,376) = 0.38$, $p = 0.536$, $\eta^2_p = 0.001$. Thus, participants who were experiencing heightened punitive motives only experienced heightened anxiety when their free will beliefs were reduced, and not when their free will beliefs were supported by a pro-free will argument nor when their punitive motives were justified for deterrence purposes.

There was no main effect for argument on anxiety, $F(2,376) = 0.28$, $p = 0.758$, $\eta^2_p = 0.001$. There was a small, marginal main effect for behavior condition: participants who read about the criminal behavior reported marginally more anxiety ($M = 3.01$, $SD = 1.15$) than participants who read about the neutral behavior ($M = 2.75$, $SD = 1.23$), $F(1,376) = 3.19$, $p = 0.075$, $\eta^2_p = 0.008$. This result is irrelevant to our hypothesis but consistent with the results of Studies 1 and 3 and prior work demonstrating that punitiveness is often associated with heightened anxiety.

We also analyzed the correlations between participants' post-argument desire to punish and anxiety. We found that greater desire to punish was unrelated to anxiety levels among participants in the Pro-Free Will, $r = 0.11$, $p = 0.187$, and the "Punish for Deterrence", $r = 0.02$, $p = 0.806$, conditions, but it was related to heightened anxiety among participants in the Anti-Free Will condition, $r = 0.17$, $p = 0.037$.

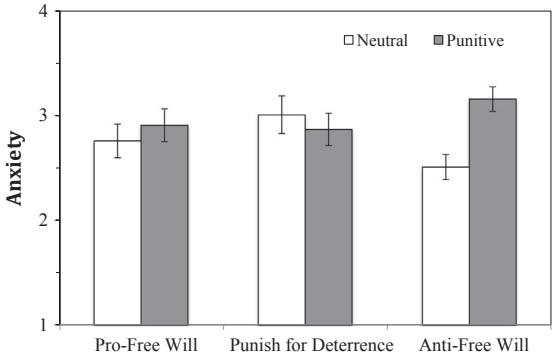

Fig. 5. Anxiety by behavior and argument conditions in Study 5.

6.3. Discussion

Study 5 confirmed further that believing in free will reduces distress over punitiveness. As in the previous studies, people who desired to punish someone, but disbelieved in free will, reported heightened anxiety. It was not the case that having one's free will beliefs challenged consistently increased anxiety, nor that desiring to punish a criminal consistently increased anxiety. Rather, the combination of exposure to a punishable offense and having one's free will beliefs undermined increased anxiety. When free will beliefs were bolstered, punitive desires did not lead to heightened anxiety.

The reasoning behind our hypothesis was that believing in a transgressor's free will justifies punishing him or her. Study 5's "Punish for Deterrence" argument provided further evidence in support of that reasoning. Participants in that condition were provided with an alternative justification for punishment (deterrence) that was explicitly independent of the existence of free will. Like the pro-free will argument, this argument removed the anxiety over punishment. Thus, it is the prospect of *unjustifiably* punishing someone that elevates distress. The present hypothesis is that belief in free will is one means of justifying punishment, but these results suggest deterrence may also justify punishment.

In fact, participants showed a significant drop in attributing free will to the transgressor after reading the "Punish for Deterrence" argument (as compared to their ratings before reading it), despite that argument said nothing about the existence of free will or lack thereof. The implication is that people attribute free will partly in order to justify punishment, and so when they are provided with an alternative justification, they no longer need to insist that the transgressor had free will. Thus, it provides further evidence that belief in free will is sometimes motivated by the wish to *justifiably* punish others.

7. General discussion

Five studies with a combined 6189 participants provided evidence that punishing others produces distress—but that people reduce that distress by believing in free will. Believing transgressors could have chosen otherwise apparently justifies punishing him or her, thereby reducing the anxiety that would otherwise be provoked by inflicting punitive harm on another person.

The present investigation used multiple methods to test the hypothesis. Studies 3–5 were laboratory experiments that manipulated beliefs about free will. Study 3 looked at spontaneous retaliation for unfair treatment. Contrary to views that revenge is sweet and satisfying, it found that people who chose to punish the unfair person felt worse than those who declined to punish him. However, these bad feelings were reduced when participants believed that the person had freely chosen to be unfair. In contrast, when participants believed the unfairness resulted from a computer glitch, punishers felt quite bad.

Studies 4 and 5 instilled the desire to punish by having participants read about a criminal transgression. Participants reported elevated levels of anxiety, but only when their belief in free will had been undermined. Those who read an article that supported free will were less bothered by their own punitive desires. Study 5 also confirmed that the anxiety was linked to a combination of low belief in free will and a high urge to punish. Only participants who had their free will beliefs reduced experienced greater anxiety after desiring to punish a criminal. In the other argument conditions, participants experiencing punitive desires were no more anxious than those not.

The assumption that belief in free will is partly motivated by desires to alleviate the distress of punishing was further supported in Study 5. Some participants were exposed to an alternative justification for punishment, based on the efficacy of deterrence and explicitly stating that free will was irrelevant. Not only did these participants show little anxiety about punishing a transgressor—they even reduced their attributions of the criminal's free will after reading the alternative justification. This suggests that when punishment is justified in utilitarian terms, there is less need to perceive transgressors as responsible for their actions.

We also reported two non-laboratory studies that sought to place these phenomena in a broader context. Using individual difference measures, Study 1 found that recent anxiety symptoms were predicted by an interaction between stable belief in free will and punitive attitudes. Believing in free will reduced the tendency for one's own punitive attitudes to predict anxiety symptoms. Using survey data and public statistics, Study 2 found that states with high levels of incarceration also had high levels of psychological distress—but only in states with citizens relatively skeptical about free will. Thus, outside the laboratory as well as in it, believing in free will seems to reduce the distress elicited by punishing others.

In sum, across three levels of analysis (individual differences, societal-level, experimental), with various operationalizations of distress (state-anxiety, trait-anxiety, punitive-specific distress, and general mental health), and across multiple samples (mechanical turk workers, *yourmorals.org* respondents, undergraduates), we found the same pattern of results: higher punitiveness in combination with lower levels of free will belief predicted higher rates of psychological distress. Together, these results suggest that free will beliefs may buffer against the distressing aspects of punitiveness by justifying punitive impulses.

7.1. Justifying punishment

The cooperative benefits and cultural ubiquity of costly punishment support its crucial role in the effective functioning of human social groups (Henrich et al., 2006). It seems clear that humans have developed an innate tendency to desire retribution against norm violators (e.g., Carlsmith et al., 2002; Cushman, 2013; Fehr & Gächter, 2002), and in some situations can even find satisfaction in punishing others (e.g., Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011).

But it is also clear that punitive motivations and behavior can sometimes lead individuals to experience a variety of negative emotions and outcomes (e.g., Carlsmith et al., 2008; McCullough et al., 2001). Punishment causes harm to the blamed and punished, and therefore must be warranted (Malle et al., 2014). Moral condemnation of people who harm others without warrant is central to

all major treatments of moral judgment (e.g., Graham et al., 2011; Gray, Young, & Waytz, 2012) and people experience guilt when they believe that their own behavior may have caused undeserved harm to another (e.g., Baumeister et al., 1994; Tangney, 1992). Anecdotal evidence of people's ambivalence toward punishment is revealed in persistent controversies over the appropriateness of severe forms of punishment (e.g., torture, capital punishment, solitary confinement); in public concerns about legal cases where punishment may have been directed toward someone not fully (or at all) responsible for the punishable offense (e.g., Campaign for Youth Justice, The Innocence Project); as well as in recent public interest in the podcast, *Serial*, and the web series, *Making a Murderer,* both non-fiction stories about potentially innocent or incompetent convicts.

It is critical to note that we do not deny that punishment can have positive affective consequences. The crucial factor that seems to tip the emotional scales regarding punishment is the degree to which the wrongdoer is perceived as personally responsible for the wrongdoing, and thus whether the punishment is perceived as justified. Our core argument is that because belief in free will facilitates moral responsibility, enhancing one's belief in free will is one means by which people are able to view norm violators as deserving of punishment, which in turn alleviates the distress that would otherwise accompany harm to a fellow human being. That the vast majority of people believe strongly in human free will (e.g., Nichols, 2004; Sarkissian et al., 2010) indicates that people already view fellow humans as generally responsible for their behavior, and thus are already equipped with the capacity to justifiably punish wrondoers. Building on work demonstrating that people bolster their belief in free will after exposure to others' immoral actions (Clark et al., 2014), we suggest that motivation to avoid the emotional upset that accompanies unjustified punishment may underlie the strength and ubiquity of such beliefs. Under our view, then, the fact that nearly all ordinary people believe very strongly in free will (despite scientific challenges and philosophical controversy) simultaneously reflects the need to justify punishment and explains why it often appears that punishing competent people needs no justification (they are believed to have free will).

However, both the present work and past work seem to suggest that bolstering one's belief in free will is not the only available means of justifying punishment. Consider research by Funk et al. (2014), which demonstrated that people often feel satisfied when unfair others are punished, specifically in cases where the punished other stated that they would change their behavior in the future. Consistent with the results of Study 5, it appears that the deterrent efficacy of punishment provides a similar consolation as assigning free will to perpetrators, namely, that the punishment was justified.

7.2. Future work and remaining questions

Future research should explore the various justifications for punishment, how these justifications interact with one another, and how these affect well-being. For example, the results of Study 5 demonstrated that reminders of the deterrent efficacy of punishment eliminated motivated increases in free will belief and reduced anxiety, suggesting the possibility that free will belief becomes less necessary when higher-level justifications are available. In other words, belief in free will may be necessary to alleviate punitive distress, but not if another rational justification is available such as that punishment is a useful way to modify behavior. It also would be interesting to see how punishment under various conditions affects the reward centers of the brain. As mentioned earlier, the anticipation of punishing a norm violator can activate a region of the brain associated with reward and pleasure (de Quervain et al., 2004). To our knowledge, it is unknown how the brain responds to actual punishment, let alone actual punishment under various conditions of justifiability.

It also remains unknown how the present findings would apply to individuals working in the criminal justice system. Despite that punishment is socially-sanctioned within these systems, police and jurors (among others) often seem to experience distress from their involvement in the punitive process. Perhaps two opposing hypotheses could be made. Because these individuals are obligated by their position to punish, and therefore the punishment is justified, free will belief may do little to further assuage their distress. Alternatively, because these individuals are not emotionally compelled to punish and likely have little anger toward the harmdoer, they may be especially in need of a justification for their involvement. Future work may seek to test these alternate possibilities.

Future work may also seek to determine how punitive attitudes and beliefs about the necessary features of moral responsibility are changing over time. In the future, shifts in our collective knowledge of external influences on human behavior could reduce retributive punitiveness and increase support for more utilitarian forms (e.g., Shariff et al., 2014). Perhaps, this would diminish the necessity of free will belief to justify punishing others, which in turn, could further diminish retributive punitiveness. Alternatively, people may come to learn that the harsh moral judgment that attends retributivism also effectively deters harmful behavior, and thus retributive attitudes may be justified even in the absence of free will. It is also possible that retributivism will decline as general humanitarianism increases regardless of how free will beliefs change over time. This raises an important question for future work. Possibly, people in modern societies feel particularly compelled to provide rational justifications for their attitudes and behaviors and free will belief is one such justification for punitive attitudes and behaviors. By this interpretation, our results are unique to the kinds of cultural systems that prize reason and rational justifications, and also consider free will a prerequisite for moral responsibility and punishment. On the other hand, our results may suggest that humans have evolved a propensity for believing in free will because it allows people to overcome harm aversion so they can ruthlessly punish those who might threaten their well-being. In other words, people who believed in free will may have been more willing to punish others retributively and this willingness deterred others from injuring them, thus increasing their evolutionary fitness. Possibly both are true, but perhaps the present results are reasonably compelling for the former interpretation and only consistent with the latter. To make a more compelling case for the latter possibility, perhaps an anthropological analysis of the emergence and history of retributive punishment and its efficacy and free will belief would be helpful.

Last, one might consider different processes for the present results in the correlational studies vs. the experimental studies. In experimental Studies 3–5, increasing free will beliefs buffered against punitive distress. Furthermore, the results of Study 5 suggested

that people actually boost their own free will belief as a means of alleviating punitive distress (at least temporarily). In correlational Studies 1 and 2, higher pre-exiting belief in free will reduced distress over pre-existing punitive attitudes. Participants in these studies likely were not actively attending to their own punitive desires while reporting their free will beliefs or distress levels. We believe it is likely that people do bolster their belief in free will as a means of reducing punitive distress, but that trait-level free will beliefs are also beneficial for the same purpose. Future work might do more to tease these processes apart.

8. Conclusion

All societies rely on punishing those who break the rules and undermine the social system's capacity to function (e.g., Henrich et al., 2006). Yet inflicting harm on others (for punishment or other purposes) can be quite difficult and psychologically distressing. People must therefore find ways to make punishment more palatable. It seems punishing someone who was powerless to act otherwise is felt as especially distressing, reflecting the principle that people should mainly be punished when they deliberately, freely choose to violate the social contract by performing harmful acts.

In order to inflict punishment, therefore, it is useful to justify the punishment by believing that the transgressor freely chose to transgress. Such a belief helps reduce the distress associated with inflicting punishment. The present investigation supported that line of work in the context of beliefs about free will. Participants felt bad about punishing others—but mainly when they thought the transgressor had been unable to act otherwise, whether because of a computer glitch or, more broadly, because of a lack of free will.

Our investigation remains agnostic about the reality of free will, which has been the focus of millennia of debate that continues even today. Despite such debates and controversies, most ordinary people maintain high belief in free will. The present findings suggest one reason for this pervasive belief in an elusive, metaphysically and scientifically mysterious phenomenon: it permits the punishment of societal menaces with diminished negative consequences for those required to observe or deliver it.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.concog.2017.03.010.

References

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

Antonio, M. E. (2006). "I didn't know it'd be so hard": Jurors' emotional reactions to serving on a capital trial. *Judicature*, *89*, 282–288.

Arendt, H. (1963). *Eichmann in Jerusalem*. New York, NY: Penguin.

Baumeister, R. F., & Brewer, L. E. (2012). Believing versus disbelieving in free will: Correlates and consequences. *Social and Personality Psychology Compass*, *6*, 736–745.

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

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*, 243–267.

Bear, A., & Bloom, P. (2016). A simple task uncovers a postdictive illusion of choice. *Psychological Science*, *27*, 914–922.

Bono, G., McCullough, M. E., & Root, L. M. (2007). Forgiveness, feeling connected to others, and well-being: Two longitudinal studies. *Personality and Social Psychology Bulletin*, *34*, 182–195.

Browning, C. R. (1993). *Ordinary men*. New York, NY: Harper Collins.

Bureau of Justice Statistics (2012a). Prisoners in 2012 - Advance Counts. Retrieved from <http://www.bjs.gov/content/pub/pdf/p12ac.pdf>.

Bureau of Justice Statistics (2012b). Probation and Parole in the United States, 2012. Retrieved from <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4844>.

Bushman, B. J. (2002). Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger, and aggressive responding. *Personality and Social Psychology Bulletin*, *28*, 724–731.

Bushman, B. J., Baumeister, R. F., & Phillips, C. M. (2001). Do people aggress to improve their mood? Catharsis beliefs, affect regulation opportunity, and aggressive responding. *Journal of Personality and Social Psychology*, *81*, 17–32.

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

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*, 1316–1324.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, *9*, 265–279.

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

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A., & Baumeister, R. F. (2014). Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology*, *16*, 501–513.

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

Cusack, R. M. (1999). *Stress and stress symptoms in capital murder jurors: Is jury duty hazardous to jurors' mental health?* (unpublished doctoral dissertation). San Antonio, TX: St. Mary's University.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*, 2–7.

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

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556.

Darley, J. M., & Zanna, M. P. (1982). Making moral judgments: Certain culturally transmitted excuses are generally believed to absolve people of blame for harming others. *American Scientist*, *70*, 515–521.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258.

Eriksson, K., Andersson, P. A., & Strimling, P. (2015). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, *19*, 152–168.

Eriksson, K., Strimling, P., & Ehn, M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. *Journal of Evolutionary Psychology*, *11*, 17–34.

Federal Bureau of Investigation (2012). Crime in the United States 2012. Retrieved from <http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/tables/4tabledatadecoverviewpdf/table_4_crime_in_the_united_states_by_region_geographic_division_and_state_2011-2012.xls>.

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

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

Funk, F., McGreer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, *40*, 986–997.

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, *45*, 840–844.
Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*, 363–374.
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130.
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*, 366–385.
Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology*, *143*, 1600–1615.
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.
Gromet, D.M., Haidt, J., & Darley, J.M. (2013). The comprehensive justice scale. Unpublished manuscript.
Grossman, D. (1996). *On killing: The psychological cost of learning to kill in war and society*. Boston, MA: Little, Brown and Company.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.
Hatzenbuehler, M. L., Keyes, K., Hamilton, A., Uddin, M., & Galea, S. (2015). The collateral damage of mass incarceration: Risk of psychiatric morbidity among nonincarcerated residents of high-incarceration neighborhoods. *American Journal of Public Health*, *105*, 138–143.
Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 451–470.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small scale societies. *Behavioral and Brain Science*, *28*, 795–855.
Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., ... Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*, 1480–1484.
Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.
Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*, 681–706.
Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, *45*, 819–828.
Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*, 78–94.
Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*, 826–842.
Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light work: The causes and consequences of social loaging. *Journal of Personality and Social Psychology*, *37*, 822–832.
Lawler-Row, K. A., Karremans, J. C., Scott, C., Edlis-Matityahou, M., & Edwards, L. (2008). Forgiveness, physiological reactivity and health: The role of anger. *International Journal of Psychophysiology*, *68*, 51–58.
Lawler-Row, K. A., & Piferi, R. L. (2006). The forgiving personality: Describing a life well lived? *Personality and Individual Differences*, *41*, 1009–1020.
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavior and Brain Sciences*, *8*, 529–566.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). *Psychological Inquiry*, *25*, 147–186.
Maltby, J., Macaskill, A., & Day, L. (2001). Failure to forgive self and others: A replication and extension of the relationship between forgiveness, personality, social desirability and general health. *Personality and Individual Differences*, *30*, 881–885.
Marteau, T. M., & Bekker, H. (1992). The development of a six-item short form of the state scale of the Spielberger State-Trait Anxiety Inventory. *British Journal of Clinical Psychology*, *31*, 301–306.
McCullough, M. E., Bellah, C. G., Kilpatrick, S. D., & Johnson, J. L. (2001). Vengefulness: Relationships with forgiveness, rumination, well-being, and the Big Five. *Personality and Social Psychology Bulletin*, *27*, 601–610.
Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*, 561–584.
National Association of State Budget Officers (2012). State Expenditure Report. Retrieved from <http://www.nasbo.org/sites/default/files/State%20Expenditure%20Report%20%28Fiscal%202012-2014%29S.pdf>.
Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language*, *19*, 473–502.
Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, *41*, 663–685.
Nietzsche, F. (1954). *Twilight of the idols* (W. Kaufmann, Trans.). New York, NY: Penguin Books (Original work published 1889).
Orbell, J., & Dawes, R. (1981). Social dilemmas. In G. Stephenson, & J. H. Davis (Vol. Eds.), *Progress in applied social psychology*. *Vol. 1*. Chichester, England: Wiley.
Orwell, G. (1943). *Looking back on the spanish war*. London: New Road.
Paulhus, D. L., & Margesson, A. (1994). Free Will and Determinism (FAD) scale. Unpublished manuscript. Vancouver, Canada: University of British Columbia.
Paulhus, D. L., & Carey, J. (2011). The FAD-Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, *93*, 96–104.
Phillips, L. H., Henry, J. D., Hosie, J. A., & Milne, A. B. (2005). Age, anger regulation and well-being. *Aging and Mental Health*, *10*, 250–256.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*, 717–731.
Roskies, A. L., & Malle, B. G. (2013). A Strawsonian look at desert. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, *16*, 133–152.
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.
Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., ... Vohs, K. D. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*, *25*, 1563–1570.
Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows present anything as significant. *Psychological Science*, *22*, 1359–1366.
Soon, C. S., Brass, M., Heinze, H., & Haynes, J. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545.
Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.
Spinaris, C. G., Denhof, M. D., & Kellaway, J. A. (2012). Posttraumatic stress disorder in United States corrections professionals: Prevalence and impact on health and functioning. *Desert Waters Correctional Outreach*, 1–32.
Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about how people may punish each other. In P. Van Lange, T. Yamagishi, & B. Rockenbach (Eds.), *Social dilemmas: Punishment and rewards* (pp. 52–69). Oxford: Oxford University Press.
Substance Abuse and Mental Health Services Administration (2012). 2011–2012 National Survey on Drug Use and Health: Model-Based Prevalence Estimates (50 States and the District of Columbia). Retrieved from <http://www.samhsa.gov/data/sites/default/files/NSDUHStateEst2011-2012/ExcelTabs/Excel/NSDUHsaeTOC2012.htm>.
Tangney, J. P. (1992). Situational determinants of shame and guilt in young adulthood. *Personality and Social Psychology Bulletin*, *18*, 199–206.
United States Department of Commerce Bureau of Economic Analysis (2012). Per Capita Real GDP by State. Retrieved from <http://www.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=1#reqid=70&step=10&isuri=1&7003=1000&7035=-1&7004=naics&7005=1&7006=xx&7036=-1&7001=11000&7002=1&7090=70&7007=2012&7093=levels>.
Vohs, K. D., & Baumeister, R. F. (2009). Addiction and free will. *Addiction Research and Theory*, *17*, 231–235.
Vohs, K. D., & Schooler, J. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*, 49–54.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences*, *7*, 65–69.
Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology*, *86*, 838–848.
Weiner, B., Graham, S., & Reyna, C. (1997). An attributional examination of retributive versus utilitarian philosophies of punishment. *Social Justice Research*, *10*, 431–452.