Moral Psychology as Accountability

Brendan de Kenessey and Stephen Darwall

In Justin D'Arms & Daniel Jacobson (eds.), *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics* (pp. 40-83). Oxford University Press, 2014. (Published under the name Brendan Dill).
Pre-publication draft. For citation or quotation, please refer to the published volume.
Comments welcome: brendan.dekenessey@gmail.com

1. Introduction

When moral psychology exploded a decade ago with groundbreaking research, there was considerable excitement about the potential fruits of collaboration between moral philosophers and moral psychologists. However, this enthusiasm soon gave way to controversy about whether either field was, or even could be, relevant to the other (e.g., Greene 2007; Berker 2009). After all, it seems at first glance that the primary question researched by moral psychologists—how people form judgments about what is morally right and wrong—is independent from the parallel question investigated by moral philosophers—what is *in fact* morally right and wrong, and why.

Once we transcend the narrow bounds of quandary ethics and "trolleyology," however, a broader look at the fields of moral psychology and moral philosophy reveals several common interests. Moral philosophers strive not only to determine what actions are morally right and wrong, but also to understand our moral concepts, practices, and psychology. They ask *what it means* to be morally right, wrong, or obligatory: what distinguishes moral principles from other norms of action, such as those of instrumental rationality, prudence, excellence, or etiquette (Anscombe 1958; Williams 1985; Gibbard 1990; Annas 1995)? Moral psychologists pursue this very question in research on the distinction between moral and conventional rules (Turiel 1983; Nichols 2002; Kelly et al. 2007; Royzman, Leeman, and Baron 2009) and in attempts to define the moral domain (e.g., Haidt and Kesebir 2010). Moral psychologists also research the question of what *motivates* moral behavior (e.g., Batson 2008), a question that philosophers have been debating from Plato's story of Gyges' ring (Plato *c.*380 BC/1992) to the present day (Nagel 1970; Foot 1972; Stocker 1976; Herman 1981; Gauthier 1986; Korsgaard 1986; Smith 1994; Shafer-Landau 2003; Rosati 2006). And right in step with the "affect revolution" in moral psychology (Haidt 2003), there has been burgeoning philosophical interest in the nature and significance of the moral emotions (Strawson 1968; Taylor 1985; Watson 1987; Wallace 1994; D'Arms and Jacobson 2000; Velleman 2001; Hieronymi 2004; Sher 2006; Scanlon 2008; Hurley and Macnamara 2010; Bagnoli 2011; Bell 2013; Coates and Tognazzini 2013).

These three areas of common interest—moral motivation, the moral emotions, and the distinctiveness of morality—present an underappreciated opportunity for fruitful collaboration between moral philosophers and moral psychologists. As a step in this direction, this chapter argues that a recent philosophical proposal regarding the nature of morality provides significant insights into moral psychology. The philosophical framework we present makes substantive psychological predictions which are well confirmed by the empirical literature.

Our central claim is that the psychology of morality, especially that of the moral motives and emotions, is best understood within a conception of morality—*morality as accountability*—that one of us has defended (Darwall 2006; 2013a; 2013b). According to morality as accountability, moral right and wrong involve accountability conceptually.[1] In taking an action to be morally wrong, for example, we take it to be something we are justifiably *held* accountable for doing through distinctive emotions and attitudes that P. F. Strawson called "reactive attitudes," such as condemnation or indignation when the agent is someone else, or guilt when it is we ourselves (Strawson 1968). Strawson argued convincingly that reactive attitudes like moral blame involve a distinctive way of regarding people, whether others or ourselves, that *holds them accountable* for complying with demands we take to be legitimate. Darwall (2006) explores the distinctively interpersonal or "second-personal" structure of these attitudes and argues that the central moral concepts of moral obligation or duty, right, and wrong are all essentially tied to second-personal attitudes and practices.

The philosophical theory of morality as accountability can be seen as a response to the question of *what it is* for an action to be morally right or wrong. Questions of the moral rightness or wrongness of an action must be carefully distinguished from other normative, ethical, and even moral questions. For example, the question of what is the most desirable life for a human being or what makes for human well-being is clearly a normative issue. But even if a good human life were to include morally right conduct as an essential element, the normative question of what makes a human life desirable is not *itself* a question of moral right and wrong. Neither is the question of what is estimable, or worthy of emulation, pride, and praise a moral issue in this sense. Many estimable accomplishments and traits are distinctively moral, of course, but many have relatively little to do with morality—artistic achievements, for example. And even if moral virtue and conduct were the only thing worthy of admiration, meriting admiration or contempt would nonetheless remain a different thing from being morally obligatory or prohibited.

Setting to one side these other morally relevant concepts and questions, let us focus on the "deontic" concepts of moral obligation or duty, right, and wrong. These concepts are interdefinable. If something is a moral obligation or duty, it follows that it would be morally wrong not to do it, and conversely. When people say that something would be "morally right," they can mean one of two things: either that it is morally obligatory or that it is not wrong, hence not obligatory not to do.

[1] We are not the first to emphasize the connection between morality and accountability. In psychology, this connection has been extensively explored in the research of Philip Tetlock (e.g. Tetlock 2002), as well as in Jonathan Haidt's recent book (Haidt 2012). In philosophy, this connection has been previously explored in the work of Allan Gibbard (1990), Gary Watson (1987, 1996), and R. Jay Wallace (1994), among others. Unfortunately, we do not have the space to survey this important work here. Our aim is to add to this literature by tying these philosophical and psychological threads together, specifically by applying the philosophical account offered in Darwall (2006) to the psychological question of the contents of the moral conscience and condemnation motives. Thanks to Jonathan Haidt for pressing us to clarify this point.

We are interested in *what it is* for an action to be morally obligatory. Since one way of referring to an action as a moral duty or obligation is to say that it is something one "morally ought" to do, and since a popular philosophical approach to "oughts" in general is in terms of *normative reasons* favoring something, a natural thought is that what it is for an action to be a moral duty is for there to be moral reason to do it.

Following a line of argument similar to G. E. Moore's famous "open question argument," however, we can see why, however plausible this may seem on theoretical grounds, it cannot be correct (Moore 1903/1993). It seems clear that two people can be completely agreed that the balance of moral reasons most favors an action, but nonetheless coherently disagree about whether the action is morally required or obligatory, that is, whether it would be morally wrong to fail to perform it. For example, an act utilitarian might hold that it would be wrong not to perform a certain action that, though it would require a massive sacrifice for the agent (say, significant harm and the risk of death), would produce greater good overall than any other act available to her in the circumstances. A critic might agree with the act utilitarian that the act would be the morally best thing the agent could do, but nonetheless hold that morality does not *require* an agent in such a circumstance to produce such an outcome at such a personal cost. Such a sacrifice, the critic might hold, would be an unreasonable one for the agent to be asked to bear to produce the outcome and, consequently, that act utilitarianism "demands too much" to be a plausible theory of moral duty, even if it were the correct theory of moral *reasons*. There can be, the critic holds, supererogatory acts that are above and beyond the call of moral duty, whereas the act utilitarian holds there to be no such category, thinking that our moral duty is always to do whatever will produce the best consequences overall.[2]

Regardless of who, act utilitarian or critic, would be right about this, it seems obvious that the two could coherently disagree about this question. But if that is so, then our interlocutors cannot both mean by "morally obligatory" being what there is most moral reason to do, since they could be agreed that the self-sacrificial, optimific act is what there is most moral reason to do in the situation but disagreed about whether this act is morally obligatory, or wrong not to do.

Darwall (2006) argues that the element that is lacking in the idea of what morality favors but essential to the deontic ideas of moral duty, obligation, or requirement, and moral prohibition or wrong, is that of moral *demands* with which we are *accountable* for complying. It is no part whatsoever of the idea that there is good reason, even good or best *moral* reason, to do something that the agent is in any way answerable for failing to do it and justifiably *held* answerable with reactive attitudes such as moral blame. This is, however, essential to the deontic ideas of moral obligation, right, and wrong. It is a conceptual truth that conduct is morally wrong if, and only if, the agent would justifiably be regarded with the attitude of moral blame by any agent (including himself), were he to undertake it without excuse. What our act utilitarian and critic were disagreeing about is whether the agent in question would be *blameworthy* if she were to fail to undertake the self-sacrificial but morally optimific action without excuse.

[2] Scheffler (1982) discusses this debate at length, ultimately siding with our critic.

We turn now to the notion of accountability and to the role and character of reactive attitudes like moral blame in holding people accountable. A central theme of Strawson's (and of Darwall 2006) is that reactive attitudes like moral blame involve a way of regarding someone that is distinctively "inter-personal" or "second personal."[3] Reactive attitudes differ from other critical attitudes, like contempt or disdain, in implicitly making a (putatively) legitimate demand of their objects and holding their objects accountable for complying with it (Strawson 1968, 85; see also Watson 1987 and Wallace 1994; for a dissenting view, see Macnamara 2013). Critical attitudes like contempt and disdain also apply an implicit standard, against which their objects are found wanting, but they do not implicitly summon their objects to hold themselves accountable for complying with this standard in the same way that reactive attitudes do. Unlike these other critical attitudes, moral blame comes with an implicit RSVP, an implicit demand for accountability and acknowledgment of the legitimacy of this demand. Whereas shame internalizes contempt, guilt *reciprocates* blame by acknowledging the legitimacy of its implicit demand, and it is itself a form of holding oneself accountable for complying with it.

It is through this analysis of the moral emotions that our philosophical proposal makes direct contact with moral psychology. Contained in the second-personal analysis of guilt and blame are specific psychological hypotheses about the cognitive presuppositions and motivational consequences of these emotions. We shall now turn to the psychological theses that we will defend throughout the remainder of this chapter, explaining how they follow from the philosophical framework laid out above.

Following a model of emotion that is popular among both psychologists (Baumeister et al. 2007) and philosophers (Hurley and Macnamara 2010), we take complex, conscious emotions such as guilt and blame to have cognitive and motivational components in addition to their basic phenomenological and physiological "feel." The cognitive component of an emotion is the state of affairs the emotion represents as obtaining. The motivational component of an emotion is the goal or motive that the emotion activates. A clear example of both these components is provided by fear, which portrays its object as dangerous (cognitive component) and motivates the subject to avoid that danger (motivational component).

The second-personal analyses of guilt and blame imply hypotheses about the cognitive and motivational components of these two emotions. Blame portrays its object (some other person) as having committed a moral wrong without excuse. And guilt makes the same "charge" concerning *oneself*. These are the cognitive components of blame and guilt, respectively. We can characterize the motivational component of both emotions as the desire to *hold the wrongdoer accountable*, whether that person is oneself (in the case of guilt) or someone else (in the case of blame). Let us elaborate on this claim. To hold someone accountable to an obligation just is to make a (putatively legitimate) moral demand of that person. When a moral wrong has already been committed, one can hold a perpetrator accountable by *pressing* the demand that was flouted via expressions of blame and

[3] "Inter-personal" is Strawson's term. Darwall (2006) uses "second-personal" to emphasize that the logical structure of these attitudes mimics the grammatical second person: they have an implicit *addressee*.

reproach. The implicit goal of these condemnatory actions is to get the perpetrator to hold *himself* accountable.

How can a perpetrator hold himself accountable? By regarding his actions as condemnable in the same way that an outside party would, and responding appropriately to this fact. He can *take* responsibility by acknowledging and internalizing the wrongness of his action. This process of holding oneself accountable always requires (i) accepting that one did wrong and is blameworthy, (ii) not merely believing that, but also having the attitude of self-blame or guilt, and (iii) internalizing the standard that one's wrongful action violated, and thus being motivated to comply with this standard in the future. Since this internalization will also motivate one to *counteract* the wrong that was done, holding oneself accountable will also often involve (iv) acknowledging one's guilt to others; (v) taking steps to ensure one's own future compliance with the violated standard; (vi) taking steps to demonstrate one's intention to comply with said standard to others; (vii) accepting punishment or sanction for one's wrongdoing; and (viii) compensating and making amends with the victims of one's wrongdoing, if there are any. By performing some or all of these actions, the perpetrator holds himself to the very demand that he had previously shirked.[4]

We can now formulate our claims about the motivational components of guilt and blame more precisely. Guilt motivates its subject to hold herself accountable by making the very demand of herself that she flouted in doing wrong; adequately holding oneself accountable will involve performing some or all of the actions listed in the previous paragraph. Blame motivates its subject to get the wrongdoer to hold himself accountable. People pursue this motive by holding the perpetrator accountable themselves, pressing the violated demand with verbal reproach, expressions of outrage, and punishment.

The motives that accompany blame and guilt are, we claim, the fundamental motives driving moral behavior. Following DeScioli and Kurzban (2009), we distinguish two primary moral motives: *conscience*, the motive to regulate one's own behavior by moral norms, and *condemnation*, the motive to respond to others' moral wrongdoing with behaviors such as reproach and punishment. Our claim is that the motivational components of blame and guilt, as described above, are also the motives of moral condemnation and conscience, respectively.

Regarding condemnation, this is a straightforward claim: the motive that accompanies blame, i.e. the motive to get the wrongdoer to hold herself accountable, is the drive behind condemnatory behavior. Regarding conscience, it is less obvious what our claim amounts to. This is because guilt is a *backward-looking* emotion, responding to a wrong action already performed, while conscience is more intuitively conceived as a *forward-looking* motivation to avoid doing wrong in one's future actions.

However, the conscience motive and the condemnation motive can each arise in both backward-looking and forward-looking contexts. The backward-looking

[4] Note that holding oneself accountable sometimes involves actions that require others' participation: e.g. (iv), (vi), (vii), and (viii) on this list. This means that holding oneself accountable is not necessarily something one can do all by oneself.

condemnation and conscience motives are, respectively, the motivational components of blame and guilt, driving one to respond to wrongdoing by holding the perpetrator accountable to the demand that was already violated (whether that perpetrator is oneself or someone else). In parallel, the forward-looking motives of conscience and condemnation aim to hold people to moral demands *before* they are violated. In forward-looking contexts, the condemnation motive is the goal to hold others to the demands of morality in their future conduct, while the conscience motive is the goal to hold *oneself* to the demands of morality in one's own future conduct.

We can now state our psychological theses in full, beginning with conscience. Looking forward, the conscience goal is to fulfill one's moral obligations, or equivalently, to comply with the moral demands to which others may legitimately hold one accountable. Looking backward, after one has already committed moral wrong without excuse, the conscience goal is to hold oneself accountable, internalizing the moral demand that one flouted, making amends, and demonstrating to others one's future intent to comply. This backward-looking conscience motive is the motivational component of the emotion of guilt, the cognitive component of which is the belief that one has committed moral wrong.

Looking forward, the condemnation goal is to hold others to the demand that they not do moral wrong. Looking backward, after another person has already committed moral wrong without excuse, the condemnation goal is to get the perpetrator to hold herself accountable for her wrongdoing. One pursues this goal by holding the perpetrator accountable oneself, pressing the demand that was flouted via verbal reproach, expressions of blame, and even punishment. This backward-looking condemnation motive is the motivational component of the emotion of blame, the cognitive component of which is the belief that some other person has committed moral wrong.

We do not of course claim that accountability-based motivations and attitudes are the only mental states that motivate moral conduct. Neither do we make any specific claims about these motives' strengths relative to other motives in human psychology. Rather, we claim only that the conscience and condemnation motives exist, have the contents we have described, and crucially, are unique in non-coincidentally motivating moral behavior. While any motive may, in Kant's phrase, "fortunately ligh[t] upon what is in fact . . . in conformity with duty" (Kant 1785/1996: 53, Ak. 4: 398), only motives and attitudes, like those we will be discussing, that involve *holding oneself to* standards of moral right and wrong can non-accidentally motivate an agent to avoid what would be morally wrong in her own view.

Our theory posits a deep unity to the moral motives. Whether the object is one's own actions or another person's, in the future or the past, the moral motives active in each of these contexts can all be described as motives to *uphold the demands of morality*. The philosophical account of morality as accountability has brought us, with some elaboration, to a unified psychological theory of the moral motives and emotions.

We will now argue that our psychological theses are supported by the existing experimental data. In section 2, we will defend our theory of condemnation; in section 3, we will defend our theory of conscience. In section 4, we will turn to the implications of our framework for the distinctive nature of morality itself.

2. Moral Condemnation

2.1. Section Prospectus

What is the motive driving morally condemnatory behavior? When we condemn, reproach, censure, sanction, or punish someone for doing wrong, what are we trying to accomplish?

Our answer to this question is that people are motivated to respond to perceived moral wrongdoing with condemnatory actions primarily in order to get the perpetrator to hold *herself* accountable for her wrongdoing. Call this the *accountability theory* of condemnation. In this section, we will argue that the accountability theory better accounts for the experimental data on condemnation than any of the available alternatives. We will begin by considering two intuitively plausible theories of condemnation, which we call the *egoistic theory* and the *deterrence theory*. Due to evidence which we shall review (§2.2 and §2.3), most psychologists have rejected both of these theories. We will then turn to the theory of condemnation that seems to be currently most popular, which we call the *retributive theory* (§2.4). Directly comparing the retributive theory with our accountability theory, we shall argue that the evidence strongly favors the accountability theory. We conclude by addressing a recent challenge to the idea that genuinely *moral* condemnation exists at all (§2.5).

2.2. The Egoistic Theory

The egoistic theory says that moral condemnation is motivated by a goal to attain a self-interested benefit of some kind. Different egoistic theories posit different self-interested motives underlying condemnatory behavior. The three most prominent views hold that condemners seek material benefit, reputational benefit, and mood improvement. We will consider each proposal in turn.

The material benefit theory says that condemnatory sanctions are applied as negative incentives with the aim of motivating others to provide material goods to the condemner. This was once taken to be the default explanation of punishment behavior in economic games such as the Public Goods game. However, studies of these very games have refuted this version of the egoistic theory. Subjects are willing to pay a fee to punish non-cooperative players even when doing so *guarantees* a material loss, since they will have no further interactions with the individuals whom they punish (Fehr and Gächter 2000; Fehr, Fischbacher, and Gächter 2002; Turillo et al. 2002; Fehr and Fischbacher 2004). If subjects were merely pursuing material or financial gain, they would not pay to punish in such conditions.

The theory that condemnatory behavior aims for reputational benefit has more empirical support. One study has shown that subjects will pay more of their own money to punish a free rider when their decision will be made known to others than when their punishment is anonymous (Kurzban, DeScioli, and O'Brien 2007). The proposed explanation of this result is that others will view a person more positively for punishing a non-cooperative other, and so people condemn in order to secure this social approval.[5] However, several studies have shown that people are willing to punish at cost to themselves even in totally anonymous conditions, which offer no opportunity for reputational gain or loss (Turillo et al. 2002; Fehr and Fischbacher 2004; Gächter and Herrmann 2008). The reputation-based egoistic theory cannot account for the motive to punish in these anonymous contexts.

The final egoistic theory we will consider says that people condemn wrongdoers in order to "let off steam," relieving the negative affect caused by the transgression and thus improving their mood. This hypothesis has been tested directly by Gollwitzer and Bushman in a recent paper titled "Do Victims of Injustice Punish to Improve their Mood?" (Gollwitzer and Bushman 2011). Two studies answer this question with a resounding *no*. Both studies confronted subjects with a free-riding perpetrator and gave them the opportunity to punish him or her. Some of these subjects were led to believe that regulating their negative mood would be ineffective or unnecessary. If condemnation were driven by a mood regulation goal, then subjects in this experimental condition would not be motivated to punish the perpetrator, since doing so would be pointless. (Indeed, a previous study using the same experimental design has demonstrated that non-morally motivated aggression *is* sometimes driven by a mood regulation goal; see Bushman, Baumeister, and Phillips 2001). However, Gollwitzer and Bushman found no significant difference in punishing behavior between the subjects in the experimental and control conditions. This finding tells strongly against the idea that condemnation is driven by an egoistic goal to improve one's mood.

As each of the three most plausible egoistic proposals faces powerful empirical objections, we conclude that we should reject the egoistic theory of condemnation as a whole.

2.3. The Deterrence Theory

The deterrence theory of condemnation says that condemnatory behaviors are motivated by the goal of deterring people from future immoral behavior. This theory is given some intuitive support by the fact that deterrence benefits seem to provide an appealing justification for condemnatory behavior, one which many people explicitly endorse (Ellsworth and Ross 1983; Carlsmith, Darley, and Robinson 2002; Carlsmith 2008).

[5] We venture that the best explanation for this result is that publicity made these subjects more motivated to do what they perceived to be morally right, i.e. punish the free rider, since accountability to others activates the conscience motive (as we argue in §3.2.2).

However, people's actual condemnatory behavior is insensitive to potential deterrence benefits, a robust finding that is fatal to the deterrence theory (Baron, Gowda, and Kunreuther 1993; Baron and Ritov 1993; Darley, Carlsmith, and Robinson 2000; Sunstein, Schkade, and Kahneman 2000; Carlsmith, Darley, and Robinson 2002; Carlsmith 2006; Carlsmith 2008; Carlsmith and Sood 2009; Keller et al. 2010). In most studies testing the deterrence theory, subjects are told about a crime and are asked to make a judgment regarding how severely the perpetrator of that crime should be punished. The experimenters then vary conditions according to some factor that is relevant to the deterrence benefits of the punishment, such as the publicity that the punishment will receive. The deterrence theory predicts that subjects will judge that a more severe punishment is appropriate when the potential deterrence benefit is high, and that a less severe punishment is appropriate when the potential deterrence benefit is low. The findings in all of these studies contradict this prediction: the severity of punishment subjects judge to be appropriate is simply insensitive to variations in potential deterrence benefit. In some studies, subjects even assign the same level of punishment when doing so will have *harmful* effects in addition to not having any deterrent effect (Baron, Gowda, and Kunreuther 1993; Baron and Ritov 1993). This robust pattern of insensitivity to deterrence benefits demonstrates that condemnatory behavior is not motivated by the goal to deter future wrongdoing.

2.4. The Accountability Theory vs. the Retributive Theory

We take the most compelling alternative to our proposal to be what we call the *retributive theory* of condemnation. This theory claims that condemnatory behavior is motivated by the goal to cause harm to the perpetrator in proportion with the blameworthiness of the perpetrator's wrongdoing. From the retributive perspective, the punishment imposed on the perpetrator is an end in itself, and the sole end condemners are after. In contrast, the accountability theory views punishment as a means to the end of getting the perpetrator to hold himself accountable for his wrongdoing. On our view, merely punishing the perpetrator is not sufficient to satisfy the condemnation motive—the perpetrator must also acknowledge his wrongdoing, feel remorse, make amends, etc., in order for the condemner's goal to be fulfilled.

2.4.1. Evidence for the Retributive Theory

The retributive theory has traditionally been presented as an alternative to the deterrence theory. As a result, empirical tests of the retributive theory have focused on testing it against the deterrence theory. The results of these experiments heavily favor the retributive theory. In addition to finding that the severity of punishment subjects judge appropriate is *not* sensitive to deterrence benefits, these studies have found that subjects' punishment judgments *are* sensitive to the perceived blameworthiness of the perpetrator (Baron, Gowda, and Kunreuther 1993; Baron and Ritov 1993; Darley, Carlsmith, and Robinson 2000; Sunstein, Schkade, and Kahneman 2000; Carlsmith, Darley and

Robinson. 2002; Carlsmith 2006; Carlsmith 2008; Carlsmith and Sood 2009; Keller et al. 2010). The more blameworthy the subjects deem the offense, the more severe the punishment they recommend—a pattern predicted by the retributive theory, but not by the deterrence theory.

Crucially, however, these studies do not support the retributive theory over the accountability theory, or vice versa. *Both* theories predict that people will judge more severe punishments to be appropriate for more blameworthy crimes. On our favored view, punishment is a part of the larger process of holding the perpetrator accountable. By imposing a punishment on the perpetrator, the condemner communicates to the perpetrator the blameworthiness of his offense, thus pushing him to internalize this blameworthiness, feel remorse, and hold himself accountable. By accepting the punishment, the perpetrator can censure himself and thereby demonstrate his commitment to the moral standard that he violated.[6] Imposing and accepting punishment are actions of the same kind as imposing and accepting verbal reproach—the former is simply a more severe version of the latter. So, just as more blameworthy actions warrant a harsher reproach, they also warrant more severe punishments. Hence the accountability theory predicts that more blameworthy actions will motivate more severe punishments, just as the retributive theory does. So we shall set the results confirming this prediction aside, and turn to experiments that can tell between these two theories.

2.4.2. Evidence for the Accountability Theory

The retributive and accountability theories make different predictions about the end state that, when attained, satisfies the condemnation motive. If the retributive theory is correct, the condemnation motive should be satisfied once the perpetrator has adequately suffered for his wrongdoing. If the accountability theory is correct, the condemnation motive should be satisfied when and only when the perpetrator has adequately held himself accountable for his wrongdoing. Thus the two theories make different predictions regarding cases where the perpetrator has been punished but has not held himself accountable by acknowledging his blameworthiness and displaying remorse.[7] If the condemnation motive is satisfied under these conditions, this evidence would favor the retributive theory. If the condemnation motive is not satisfied under these conditions, that evidence would favor the accountability theory. It would provide further support for the accountability theory if, in addition, the condemnation motive *is* satisfied when the perpetrator has acknowledged his blameworthiness and displayed remorse in addition to being punished.

[6] This characterization is confirmed by interview data on punishment in romantic relationships: "punishment sends a signal that something is wrong and a relationship rule has been broken; it helps to 'educate' an offending partner about the hurt partner and his or her needs . . . Notably, several participants also described the function of punishment as a 'test' for the relationship. If a punished partner responds with empathy and remorse, and does not retaliate in turn, then this is a reliable sign of commitment to the relationship" (Fitness and Peterson 2008, 262).

[7] The two theories also make different predictions for the case where the perpetrator hasn't been punished (in any usual sense) but has adequately held himself accountable (including to others).

To evaluate these predictions, we turn first to the research literature on *forgiveness*. McCullough (2001), quoting *Webster's Dictionary*, tells us that to forgive is "to give up resentment against or the desire to punish" (Forgive 1983, 720). If forgiveness is the giving up of blame, then it should typically result from the satisfaction and resultant deactivation of the condemnation motive.[8] This means that the accountability and retributive theories' predictions regarding the satisfaction conditions of the condemnation motive entail predictions regarding the conditions under which forgiveness occurs. Looking at the forgiveness literature through this lens, the accountability theory's predictions are overwhelmingly supported. One of the most robust findings from this research is that forgiveness occurs when and *only* when the perpetrator adequately demonstrates remorse by acknowledging guilt, apologizing, and/or offering compensation (McCullough, Worthington, and Rachal 1997; McCullough et al. 1998; Gold and Weiner 2000; Bottom, Gibson, and Daniels 2002; de Jong, Peters, and De Cremer 2003; Schmitt et al. 2004; Zechmeister et al. 2004; Kelley and Waldron 2005; Bachman and Guerrero 2006; McCullough et al. 2009; Fehr, Gelfand, and Nag 2010; Hannon et al. 2010; Leonard, Mackie, and Smith 2011; Tabak et al. 2011). In a meta-analysis of over 175 studies, Fehr, Gelfand, and Nag (2010) found that the extent to which the perpetrator apologizes predicts the degree of the victim's forgiveness with a total correlation coefficient of r = .42. This was one of the strongest effects they found. The hypothesis implied by the retributive theory, that forgiveness is predicted by the severity of the punishment received by the offender, did not have enough support to even make it onto the list of 22 predictive factors tested in Fehr et al.'s meta-analysis.

A closer look at the studies demonstrating the effects of apology on forgiveness shows that apologies only cause forgiveness when they are perceived as a sincere expression of the perpetrator's remorse and commitment to improved conduct. A mere apology is less effective than an apology combined with substantive compensation (Bottom, Gibson, and Daniels 2002). In fact, without adequate amends, an apology can backfire, resulting in *less* forgiveness (Zechmeister et al. 2004). A study that pulled apart the various elements of an apology showed that forgiveness was most likely when the perpetrator admitted fault, admitted the damage that was done, expressed remorse, and offered compensation; *only then* could the perpetrator ask for forgiveness without this request backfiring (Schmitt et al. 2004). These studies show that apologies lead to forgiveness when and only when they are seen as demonstrating that the perpetrator has fully held himself accountable for his wrongdoing. Given the premise that forgiveness is usually caused by the satisfaction of the condemnation motive, this implies that the condemnation goal aims to get the perpetrator to hold himself accountable.

Beyond the forgiveness literature, we find even more direct experimental tests of the retributive and accountability theories' predictions. A recent study titled "The Paradoxical Consequences of Revenge" reports the following surprising result: though people expect to feel better after they have punished someone who has wronged them, they in fact feel *worse* than they would if they had not punished at all (Carlsmith, Wilson,

[8] Since it is possible to voluntarily forgive a perpetrator who has neither been punished nor held herself accountable, this will not always be the case. However, as with the deliberate relinquishment of other unsatisfied goals, forgiving before one's condemnation motive has been satisfied is a difficult endeavor requiring considerable self-control; we should thus expect such cases to be relatively rare.

and Gilbert 2008). Carlsmith et al. put subjects in a Public Goods game staged to have an offender who encouraged others to cooperate and then did not cooperate herself. Some subjects were given the opportunity to punish this free rider (punishment condition) while other subjects had no opportunity to punish (no-punishment condition). After doling out punishment, the punishment condition subjects had no further interactions with or communication with the perpetrator; so, crucially for our purposes, there was no opportunity for the perpetrator to take responsibility or signal remorse. After the study, subjects in the punishment condition reported significantly more *negative* affect than subjects in the no-punishment condition. In addition, subjects who had punished the free rider reported ruminating about the offender more than subjects who had not been given the opportunity to punish.

Carlsmith et al.'s discussion focuses on the implications of these results for our understanding of affective forecasting. However, we think the study has more direct relevance to the retributive theory. According to the retributive theory, merely punishing the offender should be sufficient to fulfill the condemnation goal. Research on goal pursuit in general has shown that two important signatures of goal fulfillment are *positive* affect and *inhibition* of goal-relevant concepts; in contrast, negative affect and increased accessibility of goal-relevant concepts are signals of goal *frustration* (Chartrand 2001; Förster, Liberman, and Higgins 2005; Förster, Liberman, and Friedman 2007; Liberman, Förster, and Higgins 2007; Denzler, Förster, and Liberman 2009). Therefore, the negative affect and ruminating thoughts experienced by Carlsmith et al.'s subjects after punishing strongly indicates that merely punishing the offender *did not* fulfill their motive to condemn, contrary to the retributive theory's predictions.

Two other studies report a similar pattern (Gollwitzer and Denzler 2009; Gollwitzer, Meder, and Schmitt 2010). In both studies, subjects were given the opportunity to punish a confederate who treated them unfairly. After doling out punishment, one group of subjects received a message from the offender communicating his understanding that he deserved the punishment he had received (the "understanding" condition), while the other group either received no communication or received an actively unrepentant message from the offender. A manipulation check showed that subjects perceived the "understanding" message to "not only [contain] an admittance of harm and fault, but also an expression of remorse, an apology, and, most strikingly, a compensation offer" (Gollwitzer, Meder, and Schmitt 2010, 370). In other words, the "understanding" condition subjects took their offender to be holding himself accountable.

Confirming the accountability theory's predictions, Gollwitzer et al.'s subjects showed clear signs of goal fulfillment when they received the "understanding" message, but not when they merely punished the offender. In the first study (Gollwitzer and Denzler 2009), subjects showed significantly decreased automatic accessibility of aggression-related words after receiving the "understanding" message; as we have said, this is a strong indicator of goal fulfillment. Subjects who merely punished the offender showed no such decrease in accessibility, contrary to the retributive theory's prediction. In the second study (Gollwitzer, Meder, and Schmitt 2010), subjects reported how satisfied they felt after punishing the offender. Subjects who received the "understanding" message expressed significantly greater feelings of satisfaction than subjects in the "no understanding" condition, who, crucially, were *no more satisfied than those who had not punished the offender at all. Contra* the retributive theory, mere punishment did not satisfy subjects if it was not accompanied by the offender's holding himself accountable. Confirming the accountability theory, punishment *did* satisfy subjects if, and only if, it was accompanied by a message that those subjects saw as showing that the perpetrator was taking responsibility—admitting fault, expressing remorse, and apologizing.

Empirical tests of the predictions made by the accountability and retributive theories confirm the predictions of the accountability theory while disconfirming the predictions of the retributive theory. We have also seen (in §2.2 and §2.3) that the egoistic and deterrence theories of condemnation face significant empirical challenges as well. On this basis, we claim that the accountability theory is the best-supported theory of moral condemnation currently available. This concludes our argument for the accountability theory of condemnation.

2.5. The Existence of Genuine Moral Condemnation

We conclude our discussion of moral condemnation by considering a recent challenge to the existence of genuinely *moral* condemnation offered by C. Daniel Batson and colleagues (Batson et al. 2007; Batson, Chao, and Givens 2009; O'Mara et al. 2011). In a series of studies, Batson et al. have found that people experience and express great outrage in response to moral violations when the victim is themselves, someone in their group, or someone with whom they empathically identify, but show much less outrage at moral wrongdoing when the victim is someone else, outside of their group, with whom they do not empathize (Batson et al. 2007; Batson, Chao, and Givens 2009; O'Mara et al. 2011; see also the similar results in Yzerbyt et al. 2003; Bernhard, Fischbacher, and Fehr 2006; and Gordijn et al. 2006). They conclude from this data that the outrage experienced by their subjects cannot be *moral* outrage, because moral outrage would not differentiate in this way between offenses against oneself and one's group members on the one hand, and offenses against strangers on the other.

While we do not dispute Batson et al.'s findings, we think the conclusion they draw from these findings is mistaken. Moral outrage can take both personal and impersonal forms: the personal resentment felt by a victim of wrongdoing is no less a form of moral blame than the impersonal indignation felt by a third-party bystander (Strawson 1968; Darwall 2012). The fact that people feel personal moral outrage on behalf of themselves or others with whom they empathically identify more intensely than they feel impersonal moral outrage does not impugn the moral content of either emotion. Rather, we think this result is best explained by the simple fact that stimuli must be emotionally salient to produce a strong emotional response. Wrongs committed against oneself or members of one's group are more emotionally salient, so quite understandably, they produce a more intense response of moral outrage. A similar point holds for altruistic motivation: it can both be the case that people are sometimes genuinely driven by an altruistic motive to improve another's well-being for its own sake (Batson and Shaw 1991) *and* that people are much more likely to experience this altruistic motive when another person's welfare is made emotionally salient to them by empathic perspective-taking (Batson and Shaw 1991) or identifiability (Small and Loewenstein 2003). In general, the fact that manipulations of emotional salience affect the intensity of a motive or emotion should not affect the conclusions we draw about the *content* of that motive or emotion.

We think Batson et al. are right to emphasize the importance of distinguishing between non-moral anger and genuinely moral condemnation. However, we do not think that *impersonality* is the feature that distinguishes moral outrage from non-moral anger. Instead, we propose that the distinction between moral and non-moral anger is best characterized in terms of the differing motivations that accompany these emotions. As we have argued in this section, moral anger drives its subject toward a unique goal: to make the wrongdoer hold himself accountable to the moral demand he flouted. This goal does not seem to be shared by non-moral anger, which aims instead at regulating mood (Bushman, Baumeister, and Phillips 2001), acquiring social status (Griskevicius et al. 2009; Wenzel et al. 2008), or simply inflicting harm for its own sake (Denzler, Förster, and Liberman 2009). By showing that the condemnation motive has the essentially *moral* content of holding a perpetrator to a moral demand, as opposed to the morally neutral motives of attaining egoistic benefit, deterrence, or retribution, we have provided a basis for a principled distinction between moral and non-moral anger. This is a crucial point: previous major reviews on the moral emotions have treated anger as a unitary psychological state (Haidt 2003; Hutcherson and Gross 2011), and thus have missed a distinction of fundamental importance for moral psychology.[9]

3. Moral Conscience

3.1. Section Prospectus

What is the motive driving morally conscientious behavior? When we "do the right thing," what are we trying to accomplish?

Our answer to this question is that morally conscientious behavior is driven by *moral conscience*, an intrinsic desire to comply with moral demands to which one may be legitimately held accountable, or equivalently, to comply with one's moral obligations. This is the *accountability theory* of moral conscience.

The most popular alternative theory of moral conscience is a view we call the *approval theory*. Approval theorists are skeptical about the existence of genuine moral conscience, maintaining that instead of being motivated to *in fact* comply with our moral obligations, we are motivated only to *appear as if* we are complying with our moral obligations (Batson 2008). The approval theory denies that human beings have any

[9] An illuminating exception to this trend is Lemay, Overall, and Clark (2012), whose distinction between "hurt" and "anger" closely corresponds to our own proposed distinction between moral and non-moral anger, with what they call "hurt" corresponding to what we call "moral anger," and what they call "anger" corresponding to what we call "non-moral anger." Our proposal moves one step beyond Lemay et al.'s in hypothesizing that the moral/non-moral anger distinction applies not only to the condemnatory emotions of victims, but to those of third parties as well.

intrinsic desire to fulfill their moral obligations, instead claiming that moral behavior is driven by an instrumental desire to appear moral in order to gain the egoistic benefits of good repute. The accountability theory, in contrast, holds that while people may well be motivated to appear moral and gain a good reputation, these are not the only motives driving moral behavior; in addition, human beings have an intrinsic desire to uphold their moral duties.

The approval theory is composed of two major theses. First, it claims that morally conscientious behavior is motivated by what we will call the *approval motive*, a desire to gain the moral approval and avoid the moral disapproval of one's peers. Second, the approval theory claims that the moral conscience motive *does not exist*. In contrast, the accountability theory is not similarly committed to denying the existence of the approval motive. We think it is obvious that human beings desire approval and fear disapproval, and that this motive will sometimes contribute to the production of moral behavior. Our controversial further claim is that human agents *also* desire to uphold morality for its own sake: people are sometimes motivated by genuine moral conscience. So the disagreement between the accountability and approval theories boils down to the question of whether moral conscience exists.

Some might worry that the approval theory as we have described it is a straw man opponent, since we have saddled this theory with the very strong negative claim that moral conscience does not exist. A more plausible version of the approval theory might acknowledge that the moral conscience motive exists, but claim that it plays a minor role in producing moral behavior compared with the far more powerful approval motive. (A view like this is advocated in Haidt 2012). This hybrid view presents a more formidable opponent, which our arguments do not directly address.[10] However, we think the more radical version of the approval theory we address is worth arguing against, even if it is a straw man. Just as philosophers argue against skepticism about our knowledge of the external world, not because anyone actually holds this view, but in order to find a more secure foundation for this knowledge, it is worthwhile to argue against skepticism about moral conscience even if no one actually holds this view, in order to find a more secure foundation for our theory of moral conscience. To this end, we will focus on the stronger version of the approval theory, which denies the existence of moral conscience.[11]

The accountability theory does not merely affirm the existence of moral conscience, however: it also provides a theory of the *content* of this conscience motive. A more orthodox view of moral conscience takes the idea of moral obligation as primitive, saying that what it is for an agent to represent a rule as a moral obligation is for her to include it in her internal list of moral principles. An agent determines what principles make it onto this "moral rule list" by exercising her individual faculty of moral judgment. Regarding what distinguishes the rules that an agent takes to be her *moral obligations* from

[10] However, our arguments do have some significance for the hybrid view. We argue in §3.2 that patterns of behavior that are usually ascribed to the approval motive can be explained at least as well by appeal to the conscience motive. Our arguments thus undermine any argument based in this data for the view that the approval motive is more powerful than the conscience motive, since the data may be explained by the conscience motive as well.

[11] Thanks to Jonathan Haidt for raising this concern.

any other rules, all that seems to be said is that these rules are somehow internally stamped with a "morality label" that picks them out as the rules relevant to moral conscience. On this view, the conscience goal aims to have one's behavior conform to those rules that one has internally labeled as morally obligatory.

Contrast this orthodox view of moral conscience with the accountability theory. Rather than leaving the concept of moral obligation as primitive, the accountability theory provides an analysis of what it is to represent a rule as a moral obligation. As stated in the introduction, we hold that to regard a rule as a moral obligation just is to regard it as a demand to which any agent may legitimately hold one accountable with the reactive attitude of blame if one violates the demand without excuse. Thus we can elaborate upon the content of the conscience motive as follows: the goal to comply with one's moral obligations is one and the same as the goal to comply with those demands to which one may legitimately be held accountable. As we shall see (especially in §3.2.2), this elaborated theory of the conscience motive has explanatory resources that the more orthodox view of conscience does not.

It is important to distinguish the conscience motive as the accountability theory construes it from the motive posited by the approval theorist. The approval motive can be described as the motive to comply with those demands to which one is *actually* being held accountable by others; or to avoid *incurring* others' blame and disapproval. In contrast, the conscience motive we posit is the motive to comply with those demands to which one *would legitimately* be held accountable by others; or to avoid *warranting* others' blame and disapproval. The crucial difference is between wanting to avoid being *blamed* (approval motive) and wanting to avoid being blame*worthy* (conscience motive).

The present section will be an extended argument in favor of the accountability theory over the approval theory. The second part of the section (§3.2) will be dedicated to rebutting arguments for the approval theory; the final part (§3.3) will present a positive argument for the accountability theory based in data on guilt and shame.

Before we proceed, however, we wish to mention and set aside a third view of moral conscience that, although initially appealing, is ultimately untenable. This is the idea that the moral conscience motive is a motive of altruistic compassion for others based in empathy. Though an empathy-based altruistic motive has been shown to exist and often leads to morally right actions such as helping those in need (Batson and Shaw 1991), this motive does not have the right kind of content to either count as a moral conscience motive or explain all cases of morally conscientious behavior. The goal of empathy-based altruism is to help the person for whom empathy is felt, considered independently of the morality of doing so; thus this motive has no intrinsic moral content. In line with this conceptual point, at least one study has experimentally dissociated the empathy-based altruistic motive to help a specific person from the moral conscience motive to treat people fairly (Batson et al. 1995). This study also illustrates our second point, that there are many kinds of morally conscientious behavior that are not plausibly explained by altruistic concern for others' welfare, such as people's concern for fairness, hierarchy, authority, promissory and contractual obligations, and religious taboos. Thus, though empathy-based altruism may well sometimes help to motivate moral behavior, it cannot itself *be* the moral conscience motive. What the moral conscience motive is, and whether it exists, are the questions to which we now turn.

3.2. Arguments for the Approval Theory

In this section, we will present what we take to be the two strongest arguments for the approval theory, and argue that the accountability theory has the resources to rebut both arguments.

3.2.1. First Argument: Moral Hypocrisy and Moral Licensing

The most explicit and direct arguments for the approval theory in the literature are based upon two empirical findings: the *moral hypocrisy effect* and the *moral licensing effect.* The moral hypocrisy effect is the finding that, under certain conditions, subjects will strive to appear moral without undertaking the costs of actually acting morally (Batson et al. 1997; Batson, Thompson, and Seuferling 1999; Batson and Thompson 2001; Batson, Thompson, and Chen 2002; Batson, Collins, and Powell 2006; Batson 2008). The moral licensing effect is the finding that subjects will behave less morally immediately after engaging in behavior that makes them appear moral (Monin and Miller 2001; Cain, Loewenstein, and Moore 2005; Khan and Dhar 2006; Effron, Cameron, and Monin 2009; Sachdeva, Iliev, and Medin 2009; Mazar and Zhong 2010; Kouchaki 2011; Effron, Miller, and Monin 2012; Effron, Monin, and Miller 2013; Merritt et al. 2012). Both of these findings appear to strongly favor the approval theory because they seem to show that people are more motivated to *appear* moral than they are to actually *be* moral.

Though this prima facie appearance is strong, we think that upon reflection, these two findings do not in fact support the approval theory over the accountability theory. To see why, consider the difference between the motive to appear moral *to others* and the motive to appear moral *to oneself*. If the moral licensing and hypocrisy effects showed that subjects are less motivated to act morally when they merely appear moral *to others*, this would provide unambiguous support for the approval theory. Such a result would show that the motive driving moral behavior is satisfied when social approval has been secured, even if the agent is aware that she has not fulfilled her moral obligations. This would mean that the motive driving moral behavior is a motive to secure social approval, *not* a motive to fulfill one's moral obligations.

On the other hand, if the moral licensing and hypocrisy effects show instead that subjects are less motivated to act morally only when they appear moral *to themselves*, this result is compatible with the existence of moral conscience and thus with the accountability theory. For an agent to appear moral to herself just is for it to seem to her that she has fulfilled her moral obligations. If the agent is motivated by genuine moral conscience, she will take this motive to be satisfied when it appears to her that she has fulfilled her moral obligations. As we have already noted (in §2.4.2), it is a domain-general feature of motivation that a goal is suppressed or "turned off" when it appears to the agent that the goal has been fulfilled (Förster, Liberman, and Higgins 2005; Förster Liberman, and Friedman 2007; Liberman, Förster, and Higgins 2007; Denzler, Förster, and Liberman 2009). Thus we can predict that when it appears to an agent that she has fulfilled her moral obligations, her conscience motive will be fulfilled and thus suppressed immediately thereafter. This would make room for more selfish motives to govern the agent's behavior. Thus by affirming the existence of genuine moral conscience, the accountability theory can predict and explain why, when an agent appears moral *to herself*, she will act less morally immediately afterwards.

So, it seems that whether the moral hypocrisy and moral licensing effects ground an argument for the approval theory over the accountability theory depends crucially on whether these effects involve appearing moral *to others* or appearing moral *to oneself*. The research on these effects strongly supports the latter conclusion: moral hypocrisy and moral licensing occur when and only when subjects appear moral *to themselves*, not to others. We will now review this research, beginning with moral hypocrisy.

In the moral hypocrisy paradigm, subjects are given the opportunity to assign tasks to themselves and another participant, where one task is clearly much more enjoyable than the other. These subjects are told that the fairest choice is to flip a coin, giving oneself and the other participant equal chances of receiving the better task. About half of subjects choose to flip the coin; but crucially, these subjects still overwhelmingly assign themselves the better task (90 percent; Batson et al. 1997). This is the moral hypocrisy finding: these subjects flip the coin, which makes them appear moral, but do not obey the coin flip's results when it gives them the less enjoyable task, thus avoiding the costs of actually being moral.

Despite their unfair behavior, subjects who flip the coin subsequently rate their own behavior as having been significantly *more* moral than subjects who do not flip the coin (Batson 2008). By flipping the coin, these subjects are somehow fooling *themselves* into thinking that they did the right thing. Batson and colleagues predicted that drawing subjects' attention to their own behavior would eliminate the moral hypocrisy effect by blocking this self-deception. This is what they found: when subjects in this paradigm are placed in front of a mirror (a manipulation that has been shown to increase self-awareness), those who flipped a coin gave themselves the more enjoyable task only 50 percent of the time, abiding by the flip's results (Batson, Thompson, and Seuferling 1999).

We can sum up these findings as follows: when people are not attending to their behavior, they may convince themselves that they have behaved morally when in fact they have not (moral hypocrisy). However, as soon as someone pays enough attention to her behavior to *notice* that it is immoral (the mirror effect), she is motivated to adjust her behavior to conform to her moral standards. Clearly, Batson et al.'s moral hypocrisy effect depends on subjects' appearing moral to themselves, not to others; thus it does not provide evidence that favors the approval theory over the accountability theory.

The evidence on moral licensing points in the same direction. The original moral licensing paradigm showed that subjects are more likely to make implicitly racist or sexist hypothetical hiring decisions after being given an opportunity to explicitly disagree with racist or sexist statements (Monin and Miller 2001). The affirmations of anti-racist/anti-sexist values that produced this licensing effect were performed in a written questionnaire that subjects were told was to be private and anonymous; so, these subjects should not have seen their affirmations as having any influence over their appearance to others. A follow-up study showed that the moral licensing effect emerges just as powerfully when the value-affirming survey (the experimental manipulation) and the job selection survey (the dependent measure) are administered by two different experimenters. These findings seem to establish that the moral licensing effect is not explained by subjects' establishing their "moral credentials" to others, since they could not reasonably take their anonymous value affirmations to have any effect on anyone else's opinion of them.

Instead, the moral licensing effect seems best explained by the value affirmations causing subjects to appear moral to themselves. At least two studies have found that the moral licensing effect is statistically mediated by a positive change in how morally good subjects judge themselves to be (Sachdeva et al. 2009; Kouchaki 2011). One study obtained a moral licensing effect merely by having subjects write a story about themselves that contained morally positive words, thus affirming their own moral goodness (Sachdeva et al. 2009). Privately writing such a story has no effect on one's appearance to others, so it must produce licensing by making subjects appear moral to themselves. Thus, we submit that the moral licensing effect, like the moral hypocrisy effect, is caused by subjects' appearing moral to themselves, not to others, and thus is quite compatible with the existence of moral conscience.

Thus we conclude that the approval theory is not supported over the accountability theory by the experiments demonstrating moral hypocrisy and moral licensing, as both effects are quite compatible with the existence of moral conscience. In fact, these findings can be seen as supporting the accountability theory over the approval theory, rather than vice versa. For while the accountability theory has a ready explanation for why appearing moral to oneself should decrease subsequent moral motivation (since it fulfills the moral conscience goal), the approval theory does not. We should not expect a priori that the goal to attain the moral approval of others should be satisfied when one appears moral *only* to oneself; rather, it seems an approval-seeking agent would remain vigilant until her moral credentials were public.

Batson explains the importance of securing self-approval as follows: "If I can convince myself that serving my own interests does not violate my principles, then I can honestly appear moral and so avoid detection without paying the price of actually upholding the principles" (Batson 2008, 53). This hypothesis might be able to explain why self-approval is necessary for the satisfaction of the approval motive, but would not explain why it is *sufficient*. If deceiving yourself into approving of your own actions is a means to attaining the moral approval of others, then merely attaining self-approval should not be sufficient to satisfy the approval motive, as the moral licensing studies indicate it is. The approval motive should only be satisfied once one has gained *others'* approval in addition to one's own.

More generally, we don't need to attribute such a nefariously self-manipulating motive to human beings to explain their susceptibility to the sort of self-deception

Batson’s studies have revealed. A simpler explanation for this self-deception is that the moral conscience motive is only one motive among many, including selfish motives, which compete for control over cognition and behavior. The dominant motive at any time biases attention and cognition so as to suppress alternative, incompatible motives (Shah, Friedman, and Kruglanski 2002). So if, as seems likely, selfish motives to gain money and avoid tedious, difficult tasks are originally dominant in the experimental context, they will bias attention to avoid goal-discrepant thoughts such as “it would be unfair to disobey the coin flip and give myself the better task.” Only when moral considerations are sufficiently attention-grabbing to bring the conscience motive to the fore—as Batson et al.’s mirror ensured—will these selfish cognitive biases give way to morally motivated thought and action. We thus submit that the accountability theory gives a better explanation of the moral hypocrisy and moral licensing findings than the approval theory can provide.

3.2.2. Second Argument: The Dependence of Conscience on Social Norms

A more general, and more worrisome, argument for the approval theory goes as follows. If people are motivated by genuine moral conscience, as the accountability theory claims, then we should expect their moral behavior to be best predicted by their beliefs about what is morally right and wrong. If, on the other hand, morally conscientious behavior is solely driven by the motive to gain social approval, as the approval theory claims, then we should expect people’s moral behavior to be best predicted by whether and how their actions are being judged by others. A survey of the experimental literature overwhelmingly confirms the approval theory’s prediction that moral behavior depends in this way on social context, while providing little support for the idea that explicit moral beliefs predict moral behavior. Across many studies with various methodologies, the data indicates that whether subjects are motivated to conform to a moral standard depends primarily upon whether there are other people watching and holding them to that standard. In other words, H. L. Mencken seems to have been on target when he declared that “conscience is the inner voice that warns us somebody may be looking” (Mencken 1949). And thus, the argument goes, the approval theory is confirmed, and the accountability theory falsified.

We do not dispute the empirical premises of this argument. The data, which we will review presently, clearly indicates that moral behavior is powerfully influenced by the presence or absence of actual social accountability. However, we will argue that the accountability theory can explain this data at least as well as the approval theory can.

The first line of relevant findings shows that people’s behavior in many morally relevant domains is best predicted by their perceptions of the norms of approval and disapproval that hold in their social environment. This result has been demonstrated across many domains of moral behavior, including charitable donation (Reingen 1982), pro-environmental behavior (Reno, Cialdini, and Kallgren 1993; Kallgren, Reno, and Cialdini 2000; Schultz et al. 2007; Goldstein, Cialdini, and Griskevicius 2008), contraception use (Fekadu and Kraft 2002), voting (Gerber, Green, and Larimer 2008;

Gerber and Rogers 2009), intergroup cooperation (Paluck 2009), and a wide range of criminal behaviors (e.g. vandalization, Zimbardo 1969; tax evasion, Steenbergen, McGraw, and Scholz 1992; and many others, Grasmick and Green 1980; Tittle 1980; Kahan 1997). Summarizing the results on criminal behavior, Dan Kahan observes: "the perception that one's peers will or will not disapprove exerts a much stronger influence than does the threat of a formal sanction on whether a person decides to engage in a range of common offenses—from larceny, to burglary, to drug use" (1997, 354).

These findings are corroborated by laboratory experiments on cooperation in economic games, especially the public goods game. Cooperation in these games increases dramatically when the participants are allowed to punish others for free riding by subtracting from their money (Fehr and Gächter 2000). One might think that these subjects are cooperating simply out of self-interest, to avoid losing money by incurring sanctions. However, further studies have indicated that punishment motivates cooperation by means of its expression of disapproval rather than its financial incentives. One line of studies shows that holding the free rider accountable by merely expressing indignation is more effective than material punishment in motivating cooperation (Ostrom, Walker, and Gardner 1992; Masclet et al. 2003; Noussair and Tucker 2005; Ule et al. 2009; Janssen et al. 2010; see also Orbell, Van de Kragt, and Dawes 1988). In addition, punishment *without* moral disapproval is ineffective in motivating cooperation. Fehr and Rockenbach (2003) found that when a punishment is imposed to enforce an obviously illegitimate, selfish demand, subjects will cooperate *less* than control subjects who faced no sanctions at all. These findings indicate that what motivates subjects to cooperate are not material punishments, but rather the condemnation they express.

The approval theorist will say that the above results are best explained by the fact that agents are concerned only with their peers' approval. If people were genuinely motivated by conscience, the approval theorist may reason, they would follow their moral convictions regardless of whether doing so attracts the approval or disapproval of others.

Perhaps this conditional would hold true of an ideally rational, cognitively unlimited agent motivated solely by moral conscience. But if we consider imperfect creatures like ourselves, who have highly limited cognitive capacities and must negotiate between many competing motives, we can see why moral behavior might depend on social norms *even if* it is motivated by genuine moral conscience.

The defender of moral conscience can argue that since human beings are fallible judges of moral rectitude, it makes rational sense for us to look to others for guidance as to what is morally right and wrong. Human beings are epistemically dependent upon and highly deferential to others in judgment about descriptive matters even as mundane as the relative lengths of lines on paper (Asch 1955), and so it should be no surprise that they are similarly deferential in matters of moral judgment (Berkowitz and Walker 1967). Since people's praise and blame are good indicators of their moral beliefs, doing what others tend to praise and avoiding what others tend to blame will be a good heuristic strategy for doing what is right and avoiding what is wrong. Therefore, unless the stakes are high enough to merit going beyond this heuristic and engaging in effortful deliberation to form an independent moral judgment, a genuine conscience motive will often produce conformity to the social-moral norms expressed by others' approval and disapproval.

However, this line of thought cannot on its own provide a fully satisfactory response to the approval theorist's skeptical argument. For the dependence of moral behavior on social norms cannot be fully explained in terms of moral belief. Sometimes, social accountability motivates behavior that *violates* a person's explicit moral beliefs. The most famous and disturbing demonstration of this fact is provided by Stanley Milgram's obedience experiments, in which a majority of subjects were willing to follow the experimenter's orders to deliver painful electric shocks to another participant. When asked about this situation, most people say that it is morally wrong to continue to shock the victim past the point of danger, and insist that they would defy the experimenter's orders out of moral conviction. But when this conviction is put to the test, most people will obey the experimenter to the point of torturing another person, rather than acting on their moral belief that doing so is wrong (Milgram 1974).

Further evidence that the influence of social accountability overreaches that of moral belief is provided by Kwame Anthony Appiah's fascinating recent study of moral revolutions (Appiah 2010). Appiah investigates three rapid, society-wide changes in moral behavior: the abandonment of dueling in nineteenth-century England, the abandonment of footbinding in China, and the abolition of slavery. He summarizes his observations as follows:

> Arguments against each of these practices were well known and clearly made a good deal before they came to an end . . . Whatever happened when these immoral practices ceased, it wasn't, so it seemed to me, that people were bowled over by new moral arguments. Dueling was always murderous and irrational; footbinding was always painfully crippling; slavery was always an assault on the humanity of the slave. (Appiah 2010, xii)

Instead of being abandoned on the basis of moral argument, Appiah contends that these practices were defeated by social disapproval. The adoption of dueling by working-class men made it appear vulgar and ridiculous, forcing aristocratic gentlemen to disdain the practice that once embodied their elite code of honor. Similarly, the centuries-old Chinese practice of binding women's feet suddenly disappeared once China was mocked for it on the international stage. The British labor class spearheaded the abolitionist movement because they saw slavery as an insult to the dignity of their profession, manual labor.

What ended these immoral practices were real, on-the-ground norms of social approval and disapproval. These social norms determined moral behavior independently from moral belief: people did not act on their moral beliefs that dueling and footbinding were wrong until these practices were also socially condemned. Doesn't this show that it was the fear of disapproval, not genuine moral conscience, that motivated these revolutions in moral behavior?

Not necessarily. Since the conscience motive is just one desire among many that compete for control over an agent's behavior, it will only fully govern behavior when some feature of the agent's situation makes moral considerations sufficiently salient. The accountability theory holds that moral obligations are represented as legitimate interpersonal demands enforced by warranted attitudes of blame. So, what could be better placed to make one's moral obligations salient than actually expressed interpersonal demands and blame? Being actually held to moral standards by others makes one's accountability for complying with *warranted* demands much more salient; being actually blamed by others gives one a vivid experience of one's *blameworthiness*. So, even if conscience is a desire to avoid blameworthiness rather than actual blame, and to comply with those demands that are justified rather than those that are actually made, actual blame and actual demands can spur moral behavior by means of making motivationally salient the legitimate demands to which one is subject and the blameworthiness that their violation would entail.

In short, we claim that moral behavior depends on one's actual context of social accountability because salient cues of social accountability are usually required to activate the conscience motive sufficiently for it to control behavior. There is some independent confirmation for the hypothesis that cues of social accountability automatically spur conscience into action. Two studies have found that merely presenting an image of two eyes elicits significantly more moral behavior from subjects. One study found that subjects playing the Dictator Game on a computer with stylized eyes in the background are significantly more generous than controls (Haley and Fessler 2005). The second study replicated this result in a real-life situation: subjects were given the opportunity to serve themselves freely available coffee, which they were asked to pay for in an "honesty box." When a poster displaying a pair of eyes was placed behind the coffee dispenser, subjects paid almost three times as much for their coffee as when a control poster was displayed (Bateson, Nettle, and Roberts 2006).

The empirical premises of the approval theorist's skeptical argument can all be alternatively explained by the mechanism that underlies these priming studies. Consider first the studies showing that moral behavior tracks perceptions of social approval. Over and above the non-negligible influence of social norms on moral belief, these norms will exert independent influence on moral behavior. The awareness that others will disapprove of littering, for instance, will make motivationally salient the fact that littering is wrong, and thereby activate a conscience-based desire not to litter (Reno Cialdini, and Kallgren 1993; Kallgren, Reno, and Cialdini 2000). Similarly, the blame expressed by punishments of free riding in the public goods game makes vivid the fact that free riding is blameworthy, and thereby activates a conscience-based desire to cooperate.

A similar explanation can be applied to Appiah's historical findings. Though there were well-known moral arguments against dueling, footbinding, and slavery, the fact that these practices were socially condoned made these abstract moral principles easy to forget. (The practice of eating meat may be a modern-day analogue). Furthermore, these practices were so culturally entrenched that refraining from them attracted active social disapproval. This generated an appearance of obligation to *participate in* the practice that was far more motivationally salient than any countervailing moral arguments. These practices ended when the social sanctions for non-participation broke down, as when participation in dueling was no longer a mark of high social status, and when the practices themselves came to attract social condemnation, as when China was mocked for footbinding on the international stage. The first change dissipated the powerful appearance of an obligation to participate in the practice; the second change made the moral reprehensibility of the practice salient enough to motivate disengagement from it.

Consider, finally, the Milgram experiments. Milgram's subjects were willing to undertake the aversive task of delivering shocks because of the forceful *demands* of the experimenter, to whom they were held personally accountable. Even if the subjects believed these demands to be on balance unjustified, they nonetheless were in the grip of a strong *appearance* of their being justified (cf. Gibbard 1985, 15–17). These subjects disobeyed the experimenter, however, once their accountability to the shock victim was made more salient. In the first study, the subjects who *did* disobey the experimenter did so only once the victim protested by banging on the wall. When the victim was placed in the same room as the subjects, rendering those subjects directly accountable to the victim as well as to the experimenter, obedience of the experimenter decreased by almost 40 percent, to a minority (Milgram 1974). And when a confederate subject, also ordered to deliver shocks, vocally defied the experimenter, thereby undermining the authority of his demands, subjects overwhelmingly defied the experimenter as well (36 out of 40; Milgram 1965). In sum, though subjects in the Milgram paradigm violated their own reflective moral convictions, this may be because their conscience motives were hostage to the powerful moral appearances generated by the demands of the experimenter.[12]

We therefore conclude that the dependence of morally conscientious behavior on social norms is compatible with the accountability theory's claim that such behavior is motivated by genuine moral conscience. First, social norms of approval and disapproval can serve as a heuristic guide to moral rightness and wrongness, thus influencing people's moral behavior via their moral beliefs. Second, and more importantly, being held accountable to an actual demand by one's peers automatically activates the conscience motive to comply with legitimate demands. In the absence of social accountability, moral considerations may not be salient enough for the conscience motive to overpower other amoral motives (as in the cultures Appiah studied). And when agents are held accountable to demands they reflectively believe to be unjustified, this generates a strong appearance of moral obligation, which may be more motivationally potent than reflective moral belief (as with Milgram's subjects). Thus the accountability theory can explain the data reviewed in this section, and thereby rebut the second argument for the approval theory. The demonstrated dependence of moral behavior on social accountability is compatible with both the accountability theory and the approval theory, and so does not support one over the other.

Note, however, that this data *does* support the accountability theory over accounts of moral conscience that do not essentially implicate social accountability. The orthodox view of conscience we discussed in §3.1, which takes moral norms to be represented as mere intrapersonal standards rather than essentially interpersonal demands, cannot offer

[12] For further discussion of the role of accountability in the Milgram studies, see Darwall (2006, 162–71).

the same explanation of this data that the accountability theory can. In particular, only by positing a conceptual link between moral obligations and social accountability can one predict that the conscience motive is selectively activated by cues of social accountability; and this prediction was essential to our explanation of the Milgram and Appiah findings. In other words, it is in virtue of the accountability theory's unique thesis that moral obligations are represented as legitimate interpersonal demands that this theory is able to rebut the approval theorist's arguments for skepticism about moral conscience. Insofar as one affirms the existence of moral conscience, then, one must acknowledge its essential link to social accountability.

3.3. Argument for the Accountability Theory: Guilt as Backward-Looking Moral Conscience

The major lesson of the discussion so far seems to be that the accountability and approval theories are hard to pull apart. The data that has previously been taken to support the approval theory—the moral hypocrisy and moral licensing effects, and the dependence of moral behavior on social norms—can be explained with equal adequacy by the accountability theory. So, the data reviewed so far seems equally compatible with the existence of moral conscience as construed by the accountability theory, and the brand of skepticism about moral conscience offered by the approval theory.

However, the data we have so far considered has focused exclusively on forward-looking moral behavior, where agents are concerned with avoiding *future* violations of moral norms. Prospects for answering our question look more promising if we concentrate on how agents respond *after* they have committed moral wrong. Though they are difficult to pull apart in forward-looking contexts, the accountability theory and approval theories make sharply divergent predictions regarding this *backward-looking* moral behavior.

Consider, first, the approval theory. From the perspective of the approval motive, wrongful behavior is a PR disaster to be managed. After doing wrong, approval-driven agents should seek to mitigate the negative consequences of their wrongdoing for their reputation. Strategies for accomplishing this end include distancing oneself from the wrongdoing, providing excuses, pinning the blame on someone else, or simply avoiding social attention altogether.

In contrast, the accountability theory holds that agents who have done wrong will also be driven by a moral conscience motive to hold themselves accountable for their wrongdoing.[13] This will involve *taking* responsibility for the wrongful action rather than trying to deflect responsibility to someone else. It will involve seeking out the victim of one's actions to apologize, express remorse, and make amends, perhaps by giving

[13] It is compatible with the accountability theory that people who have done wrong may simultaneously experience both the moral conscience motive to hold themselves accountable *and* the approval motive to manage the damage to their reputations. Remember: we only claim that the moral conscience motive exists, not that the approval motive doesn't.

compensation. It will also involve a re-energized vigilance against immoral behavior, driven by the recommitment to moral standards involved in holding oneself accountable for violating those standards.

Thus the moral conscience and social approval motives should produce very different patterns of behavior after an agent has committed a moral wrong. We submit that both behavioral patterns occur, and respectively accompany the emotions of guilt and shame. Guilt is the emotional reaction to wrongdoing characteristic of the moral conscience motive, and it leads to the behaviors involved in holding oneself accountable. Shame is the emotional reaction to wrongdoing characteristic of the social approval motive, and it leads to the behaviors involved in managing one's reputation. These claims are confirmed by empirical research on guilt and shame.

The literature on guilt and shame is enormous, so we will simply state the most common findings without detailing the evidence behind them (for a helpful review, see Tangney, Stuewig, and Mashek 2007). We begin with guilt. First and foremost, guilt leads its subject to *take responsibility* for her wrongdoing (McGraw 1987; Tangney et al. 1992; Roseman, Wiest, and Swartz 1994; Baumeister, Stillwell, and Heatherton 1995; Mandel and Dhami 2005; Fisher and Exline 2006; Tracy and Robins 2006; Tangney, Stuewig, and Mashek 2007). Second, guilt motivates its subject to make amends with the victim of wrongdoing by apologizing (Roseman et al. 1994; Baumeister et al. 1995), making reparations (Roseman, Wiest, and Swartz 1994; Lickel et al. 2005; Brown et al. 2008; Zebel et al. 2008; Gino, Gu, and Zhong 2009; Čehajić-Clancy et al. 2011; de Hooge et al. 2011), striving to correct future behavior (Baumeister, Stillwell, and Heatherton 1995; Tangney et al. 1996; Amodio, Devine, and Harmon-Jones 2007; Hopfensitz and Reuben 2009; Orth, Robins, and Soto 2010; Stillman and Baumeister 2010), and even self-punishing (Bastian, Jetten, and Fasoli 2011; Nelissen 2012; Inbar et al. 2013). Guilt is characterized by other-directed empathy and concern for the victim of wrongdoing (Niedenthal, Tangney, and Gavanski 1994; Leith and Baumeister 1998; Tangney, Stuewig, and Mashek 2007; Basil, Ridgway, and Basil 2008; Yang, Yang, and Chiou 2010). Finally, guilt, whether dispositional or occurrent, leads its subject to behave more morally in general (Regan, Williams, and Sparling 1972; Cunningham, Steinberg, and Grev 1980; Montada and Schneider 1989; Tangney, Wagner, Hill-Barlow, Marschall, and Gramzow 1996; Quiles and Bybee 1997; Millar 2002; Stuewig and McCloskey 2005; Tangney, Stuewig, and Mashek 2007; Hopfensitz and Reuben 2009; Kochanska et al. 2009; Cohen et al. 2011; Polman and Ruttan 2012; for a dissenting view, see de Hooge et al. 2011).

In stark contrast with guilt, shame leads its subject to *avoid* responsibility for her wrongdoing by distancing herself from the event and blaming others for her wrongdoing (Tangney et al. 1992; Tangney, Miller, et al. 1996; Ferguson et al. 1999; Johns, Schmader, and Lickel 2005; Lickel et al. 2005). Rather than motivating apology and reconciliation, shame leads to negative interpersonal consequences such as aggression and social withdrawal (Tangney et al. 1992; Tangney, Wagner, et al. 1996; Orth et al. 2010; Cohen et al. 2011). Rather than eliciting empathy and other-directed concern, shame is associated with a focus on one's self-image and public reputation (Niedenthal et al. 1994; Tangney 1995; Smith et al. 2002; Bagozzi, Verbeke, and Gavino 2003; Lickel et al. 2005;

Tangney, Stuewig, and Mashek 2007). As shame is centered on one's overall reputation, it causes its subject to focus on her general traits; while guilt, centered on one's accountability for a particular wrongful action, draws its subject's attention to her actions rather than her overall traits (Niedenthal, Tangney, and Gavanski 1994). Finally, unlike guilt, neither dispositional nor occurrent shame has positive associations with moral behavior (Tangney, Wagner, et al. 1996; Quiles and Bybee 1997; Tangney, Stuewig, and Mashek 2007, 354; Cohen et al. 2011).

In sum, while shame motivates those behaviors we would expect to be produced by the approval motive, guilt has been shown in many empirical studies to produce exactly the behaviors we would expect to be produced by a genuine conscience motive. Thus we take the existence of a genuine backward-looking moral conscience motive, and its independence from the approval motive, to be demonstrated by the evidence showing the existence of guilt and its independence from shame.

Does this mean that we should accept the existence of forward-looking moral conscience as well? We think so. In fact, we hold that backward-looking and forward-looking moral conscience are simply manifestations of a single moral conscience motive in different contexts. The conscience goal has the same content in both contexts—to fulfill one's moral obligations and so do one's part in upholding morality— but achieving this goal requires different actions depending on whether one has already done wrong, or merely needs to avoid future wrongdoing. Thus evidence for the existence of backward-looking moral conscience is *pari passu* evidence for the existence of forward-looking moral conscience. This unity thesis regarding backward-looking and forward-looking conscience is supported by the data just reviewed showing that guilt leads to more moral behavior in general, even in areas unrelated to the guilt-inducing action (see especially Regan, Williams, and Sparling 1972). If backward-looking and forward-looking conscience are two manifestations of the same motive, then the activation of the conscience motive in a backward-looking context should also lead to greater forward-looking moral behavior. We thus conclude that the experimental evidence on guilt demonstrates the existence of moral conscience in general.

This completes our argument for the accountability theory of conscience. We have argued that genuine moral conscience exists: the objections that have motivated moral conscience skepticism are not sound (§3.2), and the empirical research on guilt and shame demonstrates the existence of the conscience motive while dissociating it from the egoistic motive to gain social approval (§3.3). Furthermore, we have argued that the content of the conscience motive should be understood in terms of social accountability. Moral obligations are represented as legitimate interpersonal demands; thus the conscience motive is a motive to comply with warranted interpersonal demands to which one may be legitimately held accountable. This unique insight of the accountability theory explains the widely confirmed observation that moral behavior depends on social context, and specifically depends upon the demands to which agents are actually held accountable by their peers (§3.2.2). Conscience motivates us to be moral, and to be moral is to be accountable.

4. What is Distinctive about Morality?

We have contended that conceiving of morality in terms of accountability provides a unified understanding of the emotions, attitudes, and motives that are targeted on acting rightly and avoiding moral wrong. The implicit goal of all moral motivation is to hold people—oneself or others—accountable for compliance with moral requirements. Moral emotions and attitudes all aim to uphold *morality*, conceived as demands with which we are accountable to one another for complying as equal moral persons. In this final section, we wish to consolidate these points by refocusing on what is distinctive about morality as an ethical conception, and, consequently, on what distinguishes the psychological items we have discussed that concern it.

It is important to emphasize again that facts about morality and moral right and wrong are *normative* and thus distinct from any descriptive psychological or social fact. Morality, in this normative sense, is what moral judgment and motivation are *about*.

In a recent article on "Morality," Haidt and Kesebir define "moral systems" as "interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make cooperative social life possible" (Haidt and Kesebir 2010, 800). By "values" and "norms," Haidt and Kesebir evidently mean psycho-social phenomena. In this sense, a value or norm consists in people *valuing* something or in their *holding* or *accepting* some norm. To hold or accept a norm or value, however, is to be committed to something normative. It is to hold that something or other is *valuable* or, for example, that some kinds of conduct are morally wrong. In this latter case, it is to hold a normative belief or attitude about morality, for example, that some kind of conduct really is morally wrong. We might think of morality, in the sense we have been discussing, as consisting in valid moral norms that make moral judgments and beliefs true or false.

Moreover, as we also noted at the outset, not all normative beliefs and attitudes concern morality. What makes for a desirable life or promotes human welfare, for example, is a normative question, but it is not, in itself, a moral issue. Even if acting morally is an essential part of well-being, the proposition that it is so is not a proposition of morality.

These observations raise two questions regarding the distinctiveness of morality. First, what distinguishes moral facts from other domains of normative fact? Second, what distinguishes the psychological attitudes, emotions, and motivations that are concerned with morality—our moral psychology—from other, non-moral normative and evaluative attitudes? Since the attitudes that make up our moral psychology are attitudes *about* morality, our answer to the first of these questions will constrain our answer to the second.

Haidt and Kesebir's definition of "moral systems" offers an answer to our second, descriptive question. However, we think that this definition fails to capture what is distinctive about moral psychology, because it fails to capture what is distinctive about morality as a normative concept. Morality is a more specific normative notion than that of just any norms and values that are concerned with "suppress[ing] or regulat[ing] selfishness and mak[ing] cooperative social life possible."

Take, for example, norms of esteem and honor that define a hierarchy of status in an honor society. We might imagine these working to regulate selfishness and foster forms of cooperation. However, the normative ideas that would be involved would be those of the honorable and the estimable, of what warrants honor, deference, and esteem, on the one hand, and contempt or disdain, on the other. These are different normative ideas from those involved in morality, as can be seen quite vividly by considering Nietzsche's critique of morality in *On the Genealogy of Morality* (Nietzsche 1887/2006). Although Nietzsche sharply criticizes morality's concepts of moral "good" and "evil," he has no objection to the concepts of "good" and "bad" of an "aristocratic" ethos that structures a hierarchy of status and honor. According to Nietzsche's etymology, the term "good" originated with the "nobles" themselves to connote qualities that fit someone for high status and to contrast with qualities that are lowly and base (Nietzsche 1887/2006, 11).

We don't have to accept Nietzsche's etymology to recognize the conceptual distinction he is marking between ideas of the noble and the base, on the one hand, and those of moral right and wrong, good and evil, on the other. This normative conceptual distinction is reflected, moreover, in the psychological difference between contempt and shame, on the one hand, and blame or condemnation and guilt, on the other. What is base or low is what warrants contempt and shame, or at least the form of shame that portrays one to oneself as contempt portrays one to the person who views one with contempt, namely, as contemptible, low, or base. Culpable wrongdoing, on the other hand, is what warrants the accountability-seeking emotions of condemnation or guilt, where guilt portrays one to oneself as condemnation portrays one to someone who condemns one, namely, as worthy of condemnation or moral blame.

In theory, an aristocratic honor code might hold to be contemptible or low the very same actions that morality condemns as morally wrong. Take Haidt and Kesebir's examples of selfishness and uncooperativeness. Morality condemns excessive selfishness or free riding on the cooperative efforts of others as morally wrong. But we can easily imagine an aristocratic ethos holding selfish free riding to be contemptible or base also. Our point is that whereas morality condemns such conduct in terms of accountability, as blameworthy lacking excuse, the normative notions involved in an honor code are fundamentally different. The attitudes and emotions they bring into play, contempt and shame, differ from the accountability-seeking attitudes implicated in morality. When a noble looks down on a serf with contempt, he hardly aims to have the serf hold himself accountable for his contemptible state. The emotion that responds to contempt is not guilt, but shame, which, as we have noted (§3.3), shows itself in very different ways than guilt does.

Moreover, just as moral norms differ conceptually from those of honor and esteem, so also are they conceptually distinct from norms of purity and disgust. It is a favorite claim of Jonathan Haidt's that liberals tend to think of morality in terms of fairness and equality, whereas conservatives also include loyalty, respect for authority, and purity among morality's "foundations" (Haidt and Kesebir 2010, 821-3). Whereas liberals are skeptical of "disgust-based" notions of purity, Haidt claims that disgust is a *moral* emotion and, therefore, that purity is no less a foundation of morality than are equality and fairness.

For purposes of discussion, we can simply stipulate that an action that violates a purity taboo, consensual incest, for example, is morally wrong, indeed that it is wrong for that very reason. Our point is that the proposition that incest is morally wrong is a different kind of normative claim than the proposition that it is impure, or violates a taboo, or that it is disgusting. For present purposes, we can even allow that there is a genuine normative concept of the disgusting—of what justifies disgust or to which disgust is a fitting response—that is distinct from the concept of what actually causes disgust. Our point remains that the proposition that it is morally wrong to do something disgusting is an additional moral claim that goes beyond the claim that such an action warrants disgust. It is the claim that the disgusting action *also* warrants the accountability-seeking attitudes of condemnation and guilt.[14]

Consideration of these examples shows that Haidt and Kesebir's functional definition of morality is too broad. Norms of honor and contempt, or of purity and disgust, count as 'moral systems' by their definition, insofar as they help to suppress selfishness and promote cooperation. However, lumping these together into a single category with moral norms of blame and guilt elides an important distinction: between norms that simply *evaluate* people, such as those of honor and purity, and those that *hold people responsible*, in the sense of asking for a *response*. It is this latter feature that, we claim, makes morality distinctive.

Thus we propose an alternative definition of morality. Morality, understood as a normative phenomenon, is that set of demands to which we may legitimately hold one another accountable with blame, and hold ourselves accountable with guilt. *Moral* attitudes, emotions, and motives are those the contents of which essentially refer to morality so understood. A particular society's morality, understood as a descriptive phenomenon, is that set of norms to which the members of the society *actually* hold one another accountable with blame, and hold themselves accountable with guilt.

We should stress that according to morality as accountability, what makes a set of normative beliefs and practices *moral* beliefs and practices is not their contents—the actions they prescribe and proscribe—but the distinctively accountability-seeking attitudes and responses they take those actions to warrant. The definitions of "morality" to which Haidt and Kesebir object are all content-based definitions, such as Turiel's definition of moral judgments as "prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other" (Turiel 1983, 3). Haidt and Kesebir's objection is that these characterizations of morality typically limit the potential contents of moral beliefs to those endorsed by Western liberals. Our account avoids this objection, since it is content independent.

[14] We can even agree that there is such a thing as moral disgust. However, it seems that any such response itself presupposes the concepts of moral right and wrong and, therefore, on our analysis, the independent idea of actions warranting the accountability-seeking emotions of condemnation and guilt.

To illustrate, notice that we can agree with Haidt and Kesebir that conservatives' beliefs and attitudes are no less moral, that is beliefs *about* morality, than are liberals', regardless of whose beliefs are more correct. What is more, we can say what the difference between the conservative and the liberal amounts to: a disagreement about what standards we can hold each other accountable to. A liberal might be just as disgusted by some behavior as a conservative, but disagree with a conservative's belief that the conduct is morally wrong, since only the latter thinks that the behavior warrants blame as well as disgust.

What is distinctive about morality as a normative idea, therefore, is neither its content nor its performing the functions of regulating social order, curbing selfishness, or enabling social cooperation in just any way. It is the specific way morality purports to regulate. Morality's distinctiveness, and its distinctive form of social regulation, are explained by its conceptual tie to mutual accountability and to the accountability-seeking attitudes of blame and guilt. We have been arguing that appreciating this fact enables a unified and explanatory psychological account of moral thought, emotion, and motivation.[15]

[15] We thank Jonathan Haidt, Joshua Knobe, and three anonymous reviewers for helpful comments and criticism.

References

Amodio, D. M., Devine, P. G., and Harmon-Jones, E. 2007. A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science*, *18*(6), 524–30.

Annas, J. 1995. Prudence and morality in ancient and modern ethics. *Ethics*, *105*(2), 241–257.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy*, *33*(124), 1–19.

Appiah, K. A. 2010. *The Honor Code: How Moral Revolutions Happen*. New York: W.W. Norton & Company.

Asch, S. E. 1955. Opinions and social pressure. *Scientific American*, *193*(4), 31–5.

Bachman, G. F., and Guerrero, L. K. 2006. Forgiveness, apology, and communicative responses to hurtful events. *Communication Reports*, *19*(1), 45–56.

Bagnoli, C. (ed.) 2011. *Morality and the Emotions*. Oxford: Oxford University Press.

Bagozzi, R. P., Verbeke, W., and Gavino, J. C. J. 2003. Culture moderates the self-regulation of shame and its effects on performance: The case of salespersons in the Netherlands and the Philippines. *Journal of Applied Psychology*, *88*(2), 219–33.

Baron, J., Gowda, R., and Kunreuther, H. 1993. Attitudes toward managing hazardous waste: What should be cleaned up and who should pay for it? *Risk Analysis*, *13*(2), 183–92.

Baron, J., and Ritov, I. 1993. Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty*, *7*(1), 17–33.

Basil, D. Z., Ridgway, N. M., and Basil, M. D. 2008. Guilt and giving: A process model of empathy and efficacy. *Psychology and Marketing*, *25*(1), 1–23.

Bastian, B., Jetten, J., and Fasoli, F. 2011. Cleansing the soul by hurting the flesh: The guilt-reducing effect of pain. *Psychological Science*, *22*(3), 334–5.

Bateson, M., Nettle, D., and Roberts, G. 2006. Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*, 412–14.

Batson, C. D. 2008. Moral masquerades: Experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences*, *7*, 51–66.

Batson, C. D., Chao, M. C., and Givens, J. M. 2009. Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, *45*(1), 155–60.

Batson, C. D., Collins, E., and Powell, A. A. 2006. Doing business after the fall: The virtue of moral hypocrisy. *Journal of Business Ethics*, *66*(4), 321–35.

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E. L., Fleming, D. A., Marzette, C. M., Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., Zerger, T. 2007. Anger at unfairness: is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272–85.

Batson, C. D., Klein, T. R., Highberger, L., and Shaw, L. L. 1995. Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology*, *68*(6), 1042–54.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J., Kampf, H., and Wilson, A. 1997. In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, *72*(6), 1335–48.

Batson, C. D., and Shaw, L. L. 1991. Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*2, 107–22.

Batson, C. D., and Thompson, E. R. 2001. Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science*, *10*2, 54–7.

Batson, C. D., Thompson, E. R., and Chen, H. 2002. Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330–9.

Batson, C. D., Thompson, E., and Seuferling, G. 1999. Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–37.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. 1995. Personal narratives about guilt: Role in action control and interpersonal relationships. *Basic and Applied Social Psychology*, *17*(1), 173–98.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., and Zhang, L. 2007. How emotion shapes behavior: feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, *11*(2), 167–203.

Bell, M. 2013. *Hard Feelings: The Moral Psychology of Contempt*. New York: Oxford University Press.

Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4): 293–329.

Berkowitz, L., and Walker, N. 1967. Laws and moral judgments. *Sociometry*, *30*(4), 410–22.

Bernhard, H., Fischbacher, U., and Fehr, E. 2006. Parochial altruism in humans. *Nature*, *442*, 912–15.

Bottom, W., Gibson, K., and Daniels, S. 2002. When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, *13*(5), 497–513.

Brown, R., González, R., Zagefka, H., Manzi, J., and Cehajic, S. 2008. Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology*, *94*(1), 75–90.

Bushman, B. J., Baumeister, R. F., and 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*(1), 17–32.

Cain, D. M., Loewenstein, G., and Moore, D. A. 2005. The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *The Journal of Legal Studies*, *34*(1), 1–25.

Carlsmith, K. M. 2006. The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–51.

Carlsmith, K. M. 2008. On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119–37.

Carlsmith, K. M., Darley, J. M., and Robinson, P. H. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*2, 284–99.

Carlsmith, K. M., and Sood, A. M. 2009. The fine line between interrogation and retribution. *Journal of Experimental Social Psychology*, *45*(1), 191–6.

Carlsmith, K. M., Wilson, T. D., and Gilbert, D. T. 2008. The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316–24.

Čehajić-Clancy, S., Effron, D. A., Halperin, E., Liberman, V., and Ross, L. D. 2011. Affirmation, acknowledgment of in-group responsibility, group-based guilt, and support for reparative measures. *Journal of Personality and Social Psychology*, *101*(2), 256–70.

Chartrand, T. L. 2001. Mystery moods and perplexing performance: Consequences of succeeding and failing at a nonconscious goal. Unpublished manuscript.

Coates, D. J., and Tognazzini, N. A. (eds.). 2013. *Blame: Its Nature and Norms*. Oxford: Oxford University Press.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. 2011. Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology*, *100*(5), 947–66.

Cunningham, M., Steinberg, J., and Grev, R. 1980. Wanting to and having to help: Separate motivations for positive mood and guilt-induced helping. *Journal of Personality and Social Psychology*, *38*(2), 181–92.

Darley, J. M., Carlsmith, K. M., and Robinson, P. H. 2000. Incapacitation and just deserts as motives for punishment. *Law and human behavior*, *24*6, 659–83.

D'Arms, J., and Jacobson, D. 2000. Sentiment and value. *Ethics* 110: 722–48.

Darwall, S. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Darwall, S. 2012. Bipolar obligation. In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*, vol. 7, pp. 333-67. Oxford: Oxford University Press.

Darwall, S. 2013a. *Morality, Authority, and Law: Essays in Second-Personal Ethics I.* Oxford: Oxford University Press.

Darwall, S. 2013b. *Honor, History, and Relationship: Essays in Second-Personal Ethics II.* Oxford: Oxford University Press.

de Hooge, I. E., Nelissen, R. M. A., Breugelmans, S. M., and Zeelenberg, M. 2011. What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology*, *100*(3), 462–73.

de Jong, P. J., Peters, M. L., and De Cremer, D. 2003. Blushing may signify guilt: Revealing effects of blushing in ambiguous social situations. *Motivation and Emotion*, *27*(3), 225–49.

Denzler, M., Förster, J., and Liberman, N. 2009. How goal-fulfillment decreases aggression. *Journal of Experimental Social Psychology*, *45*(1), 90–100.

DeScioli, P., and Kurzban, R. 2009. Mysteries of morality. *Cognition*, *112*(2), 281–99.

Effron, D. A., Cameron, J. S., and Monin, B. 2009. Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology*, *45*(3), 590–3.

Effron, D. A., Miller, D. T., and Monin, B. 2012. Inventing racist roads not taken: The licensing effect of immoral counterfactual behaviors. *Journal of Personality and Social Psychology*, *103*(6), 916–32.

Effron, D. A., Monin, B., and Miller, D. T. 2013. The unhealthy road not taken: Licensing indulgence by exaggerating counterfactual sins. *Journal of Experimental Social Psychology*, *49*3, 573–8.

Ellsworth, P. C., and Ross, L. 1983. Public opinion and capital punishment: A close examination of the views of abolitionists and retentionists. *Crime & Delinquency*, *29*(1), 116–69.

Fehr, E., and Fischbacher, U. 2004. Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–94.

Fehr, E., Fischbacher, U., and Gächter, S. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1–25.

Fehr, R., Gelfand, M. J., and Nag, M. 2010. The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin*, *136*(5), 894–914.

Fehr, E., and Rockenbach, B. 2003. Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137–40.

Fekadu, Z., and Kraft, P. 2002. Expanding the theory of planned behavior: The role of social norms and group identification. *Journal of Health Psychology*, *7*(1), 33–43.

Ferguson, T. J., Stegge, H., Miller, E. R., and Olsen, M. E. 1999. Guilt, shame, and symptoms in children. *Developmental Psychology*, *35*(2), 347–57.

Fisher, M. L., and Exline, J. J. 2006. Self-forgiveness versus excusing: The roles of remorse, effort, and acceptance of responsibility. *Self and Identity*, *5*(2), 127–46.

Fitness, J., and Peterson, J. 2008. Punishment and forgiveness in close relationships: An evolutionary, social-psychological perspective. In J. P. Forgas and J. Peterson (eds.), *Social Relationships: Cognitive, Affective, and Motivational Processes*, pp. 255–69. New York: Taylor & Francis Group, LLC.

Foot, P. 1972. Morality as a system of hypothetical imperatives. *The Philosophical Review*, *81*, 305–16.

Forgive. 1983. In: *Webster's New Universal Unabridged Dictionary.* New York: Dorset & Baker.

Förster, J., Liberman, N., and Friedman, R. S. 2007. Seven principles of goal activation: A systematic approach to distinguishing goal priming from priming of non-goal constructs. *Personality and Social Psychology Review*, *11*(3), 211–33.

Förster, J., Liberman, N., and Higgins, E. T. 2005. Accessibility from active and fulfilled goals. *Journal of Experimental Social Psychology*, *41*(3), 220–39.

Gächter, S., and Herrmann, B. 2008. Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791–806.

Gauthier, D. 1986. *Morals by Agreement.* Oxford: Oxford University Press.

Gerber, A. S., Green, D. P., and Larimer, C. W. 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. *The American Political Science Review*, *102*(1), 33–48.

Gerber, A. S., and Rogers, T. 2009. Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics, 71*(1), 178–91.

Gibbard, A. 1985. Moral judgment and the acceptance of norms. *Ethics*, *96*(1), 5–21.

Gibbard, A. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge, MA: Harvard University Press.

Gino, F., Gu, J., and Zhong, C. B. 2009. Contagion or restitution? When bad apples can motivate ethical behavior. *Journal of Experimental Social Psychology*, *45*(6), 1299–302.

Gold, G. J., and Weiner, B. 2000. Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology*, *22*(4), 291–300.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *The Journal of Consumer Research*, *35*(3), 472–82.

Gollwitzer, M., and Bushman, B. J. 2011. Do victims of injustice punish to improve their mood? *Social Psychological and Personality Science*, 00(0), 1–9.

Gollwitzer, M., and Denzler, M. 2009. What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, *45*(4), 840–4.

Gollwitzer, M., Meder, M., and Schmitt, M. 2010. What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*(3), 364–74.

Gordijn, E. H., Yzerbyt, V., Wigboldus, D., and Dumont, M. 2006. Emotional reactions to harmful intergroup behavior. *European Journal of Social Psychology*, *36*(1), 15–30.

Grasmick, H. G. and Green, D. E. 1980. Legal punishment, social disapproval and internalization as inhibitors of illegal behavior. *The Journal of Criminal Law and Criminology*, *71*(3), 325–35.

Greene, J. D. 2007. The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.), *Moral Psychology*, vol. 3: *The Neuroscience of Morality: Emotion, Disease, and Development*, pp. 359–71. Cambridge, MA: MIT Press.

Griskevicius, V., Tybur, J. M., Gangestad, S. W., Perea, E. F., Shapiro, J. R., and Kenrick, D. T. 2009. Aggress to impress: Hostility as an evolved context-dependent strategy. *Journal of Personality and Social Psychology*, *96*(5), 980–94.

Haidt, J. 2003. The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), *Handbook of Affective Sciences*, pp. 852–70. Oxford: Oxford University Press.

Haidt, J., and Kesebir, S. 2010. Morality. In S. T. Fiske, D. T. Gilbert, and G. Lindzey (eds.), *Handbook of Social Psychology*, 5th edn., pp. 797–832. Hoboken, NJ: John Wiley & Sons, Inc.

Haidt, J. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY: Random House, Inc.

Haley, K. J., and Fessler, D. M. T. 2005. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245–56.

Hannon, P. A., Rusbult, C. E., Finkel, E. J., and Kamashiro, M. 2010. In the wake of betrayal: Amends, forgiveness, and the resolution of betrayal. *Personal Relationships*, *17*(2), 253–78.

Herman, B. 1981. On the value of acting from the motive of duty. *The Philosophical Review*, *90*(3), 359–82.

Hieronymi, P. 2004. The force and fairness of blame. *Philosophical Perspectives*, *18*(1), 115–148.

Hopfensitz, A., and Reuben, E. 2009. The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, *119*(540), 1534–59.

Hurley, E. and Macnamara, C. 2010. Beyond belief: Toward a theory of the reactive attitudes. *Philosophical Papers*, *39*(3), 373–99.

Hutcherson, C. A., and Gross, J. J. 2011. The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, *100*(4), 719–37.

Inbar, Y., Pizarro, D. A., Gilovich, T., and Ariely, D. 2013. Moral masochism: On the connection between guilt and self-punishment. *Emotion*, *13*(1), 14–18.

Janssen, M. A., Holahan, R., Lee, A., and Ostrom, E. 2010. Lab experiments for the study of social-ecological systems. *Science*, *328*(5978), 613–17.

Johns, M., Schmader, T., and Lickel, B. 2005. Ashamed to be an American? The role of identification in predicting vicarious shame for anti-Arab prejudice after 9–11. *Self and Identity*, *4*(4), 331–48.

Kahan, D. M. 1997. Social influence, social meaning, and deterrence. *Virginia Law Review*, *83*(2), 349–95.

Kallgren, C. A., Reno, R. R., and Cialdini, R. B. 2000. A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, *26*, 1002.

Kant, I. 1785/1996. *Groundwork of the Metaphysics of Morals*. In *Practical Philosophy*, trans. and ed. M. J. Gregor. Cambridge: Cambridge University Press. References are to page numbers of the Preussische Akademie edition.

Keller, L. B., Oswald, M. E., Stucki, I., and Gollwitzer, M. 2010. A closer look at an eye for an eye: laypersons' punishment decisions are primarily driven by retributive motives. *Social Justice Research*, *23*(2–3), 99–116.

Kelley, D. L., and Waldron, V. R. 2005. An investigation of forgiveness-seeking communication and relational outcomes. *Communication Quarterly*, *53*(3), 339–58.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., and Fessler, D. M. T. 2007. Harm, affect, and the moral/conventional distinction. *Mind and Language*, *22*(2), 117–31.

Khan, U., and Dhar, R. 2006. Licensing effect in consumer choice. *Journal of Marketing Research*, *43*, 259–66.

Kochanska, G., Barry, R. A., Jimenez, N. B., Hollatz, A. L., and Woodard, J. 2009. Guilt and effortful control: Two mechanisms that prevent disruptive developmental trajectories. *Journal of Personality and Social Psychology*, *97*(2), 322–33.

Korsgaard, C. 1986. Skepticism about practical reason. *The Journal of Philosophy*, *83*(1), 5–25.

Kouchaki, M. 2011. Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of Personality and Social Psychology*, *101*(4), 702–15.

Kurzban, R., DeScioli, P., and O'Brien, E. 2007. Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84.

Leith, K. P., and Baumeister, R. F. 1998. Empathy, shame, guilt, and narratives of interpersonal conflicts: Guilt-prone people are better at perspective taking. *Journal of Personality*, *66*(1), 1–37.

Lemay, E. P., Overall, N. C., and Clark, M. S. 2012. Experiences and interpersonal consequences of hurt feelings and anger. *Journal of Personality and Social Psychology, 103*(6), 982–1006.

Leonard, D. J., Mackie, D. M., and Smith, E. R. 2011. Emotional responses to intergroup apology mediate intergroup forgiveness and retribution. *Journal of Experimental Social Psychology*, *47*(6), 1198–206.

Liberman, N., Förster, J., and Higgins, E. T. 2007. Completed vs. interrupted priming: reduced accessibility from post-fulfillment inhibition. *Journal of Experimental Social Psychology*, *43*(2), 258–64.

Lickel, B., Schmader, T., Curtis, M., Scarnier, M., and Ames, D. R. 2005. Vicarious shame and guilt. *Group Processes & Intergroup Relations*, *8*(2), 145–57.

McCullough, M. E. 2001. Forgiveness: Who does it and how do they do it? *Current Directions in Psychological Science*, *10*6, 194.

McCullough, M. E., Rachal, K. C., Sandage, S. J., Worthington, E. L., Brown, S. W., and Hight, T. L. 1998. Interpersonal forgiving in close relationships: II. Theoretical elaboration and measurement. *Journal of Personality and Social Psychology*, *75*6, 1586–603.

McCullough, M., Root, L., Tabak, B., and Witvliet, C. 2009. Forgiveness. In C. R. Snyder and S. J. Lopez (eds.), *Oxford Handbook of Positive Psychology*, 2nd edn., pp. 427–36. New York: Oxford University Press.

McCullough, M. E., Worthington, E. L., and Rachal, K. C. 1997. Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology*, *73*2, 321–36.

McGraw, K. M. 1987. Guilt following transgression: An attribution of responsibility approach. *Journal of Personality and Social Psychology*, *53*2, 247–56.

Macnamara, C. 2013. Taking demands out of blame. In D. J. Coates and N. A. Tognazzini (eds.), *Blame: Its Nature and Norms*, pp. 141–61. Oxford: Oxford University Press.

Mandel, D. R., and Dhami, M. K. 2005. "What I did" versus "what I might have done": Effect of factual versus counterfactual thinking on blame, guilt, and shame in prisoners. *Journal of Experimental Social Psychology*, *41*(6), 627–35.

Masclet, D., Noussair, C., Tucker, S., and Villeval, M. C. 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review*, *93*(1), 366–80.

Mazar, N., and Zhong, C. B. 2010. Do green products make us better people? *Psychological Science*, *21*(4), 494–8.

Mencken, H. L. 1949. *A Mencken Chrestomathy*. New York: Knopf.

Merritt, A. C., Effron, D. A., Fein, S., Savitsky, K. K., Tuller, D. M., and Monin, B. 2012. The strategic pursuit of moral credentials. *Journal of Experimental Social Psychology*, *48*(3), 774–7.

Milgram, S. 1965. Liberating effects of group pressure. *Journal of Personality and Social Psychology*, *1*(2), 127–34.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row Publishers, Inc.

Millar, M. 2002. Effects of a guilt induction and guilt reduction on door in the face. *Communication Research*, *29*(6), 666–80.

Monin, B., and Miller, D. T. 2001. Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33–43.

Montada, L., and Schneider, A. 1989. Justice and emotional reactions to the disadvantaged. *Social Justice Research*, *3*(4), 313–44.

Moore, G. E. 1993. *Principia Ethics*, ed. T. Baldwin. Cambridge: Cambridge University Press.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Oxford University Press.

Nelissen, R. M. A. 2012. Guilt induced self-punishment as a sign of remorse. *Social Psychological and Personality Science*, *3*(2), 139–44.

Nichols, S. 2002. Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*(2), 221–36.

Niedenthal, P. M., Tangney, J. P., and Gavanski, I. 1994. "If only I weren't" versus "if only I hadn't": distinguishing shame and guilt in counterfactual thinking. *Journal of Personality and Social Psychology*, *67*(4), 585–95.

Nietzsche, F. 1887/2006. *On the Genealogy of Morals and Other Writings: Revised Student Edition*, trans. C. Diethe, ed. K. Ansell-Pearson. Cambridge: Cambridge University Press.

Noussair, C., and Tucker, S. 2005. Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, *43*(3), 649–60.

O'Mara, E. M., Jackson, L. E., Batson, C. D., and Gaertner, L. 2011. Will moral outrage stand up? Distinguishing among emotional reactions to a moral violation. *European Journal of Social Psychology*, *41*(2), 173–9.

Orbell, J. M., Van de Kragt, A. J., and Dawes, R. M. 1988. Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology*, *54*(5), 811–19.

Orth, U., Robins, R. W., and Soto, C. J. 2010. Tracking the trajectory of shame, guilt, and pride across the life span. *Journal of Personality and Social Psychology*, *99*(6), 1061–71.

Ostrom, E., Walker, J., and Gardner, R. 1992. Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*, *86*(2), 404–17.

Paluck, E. L. 2009. Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology*, *96*(3), 574–87.

Polman, E., and Ruttan, R. L. 2012. Effects of anger, guilt, and envy on moral hypocrisy. *Personality and Social Psychology Bulletin*, *38*(1), 129–39.

Plato. 1992. *Republic*, rev. edn., trans. G. M. A. Grube and C. D. C. Reeve. Indianapolis: Hackett.

Quiles, Z. N., and Bybee, J. 1997. Chronic and predispositional guilt: Relations to mental health, prosocial behavior, and religiosity. *Journal of Personality Assessment*, *69*(1), 104–26.

Regan, D. T., Williams, M., and Sparling, S. 1972. Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology*, *24*(1), 42–5.

Reingen, P. H. 1982. Test of a list procedure for inducing compliance with a request to donate money. *Journal of Applied Psychology*, *67*(1), 110–18.

Reno, R. R., Cialdini, R. B., and Kallgren, C. A. 1993. The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–12.

Rosati, C. S. 2006. Moral motivation. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2008 edn.

Roseman, I. J., Wiest, C., and Swartz, T. S. 1994. Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology*, *67*(2), 206–21.

Royzman, E. B., Leeman, R. F., and Baron, J. 2009. Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition*, *112*(1), 159–74.

Sachdeva, S., Iliev, R., and Medin, D. L. 2009. Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*(4), 523–8.

Scanlon, T. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Scheffler, S. 1982. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford: Oxford University Press.

Schmitt, M., Gollwitzer, M., Förster, N., and Montada, L. 2004. Effects of objective and subjective account components on forgiving. *The Journal of Social Psychology*, *144*(5), 465–86.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–34.

Shafer-Landau, R. 2003. *Moral Realism: A Defence*. Oxford: Clarendon Press.

Shah, J. Y., Friedman, R., and Kruglanski, A. W. 2002. Forgetting all else: On the antecedents and consequences of goal shielding. *Journal of Personality and Social Psychology*, *83*(6), 1261–80.

Sher, G. 2006. *In Praise of Blame*. Oxford: Oxford University Press.

Small, D. A., and Loewenstein, G. 2003. Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*(1), 5–16.

Smith, M. 1994. *The Moral Problem*. Malden, MA: Blackwell Publishing Ltd.

Smith, R. H., Webster, J. M., Parrott, W. G., and Eyre, H. L. 2002. The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, *83*(1), 138–59.

Steenbergen, M. R., McGraw, K. M., and Scholz, J. T. 1992. Taxpayer adaptation to the 1986 tax reform act: Do new tax laws affect the way taxpayers think about taxes? In J. Slemrod (ed.), *Why People Pay Taxes: Tax Compliance and Enforcement*. Ann Arbor: University of Michigan Press.

Stillman, T. F., and Baumeister, R. F. 2010. Guilty, free, and wise: Determinism and psychopathy diminish learning from negative emotions. *Journal of Experimental Social Psychology*, *46*(6), 951–60.

Stocker, M. 1976. The schizophrenia of modern ethical theories. *The Journal of Philosophy*, *73*(14), 453–66.

Strawson, P. F. 1968. Freedom and resentment. In P. F. Strawson and G. Ryle (eds.), *Studies in the Philosophy of Thought and Action*, pp. 71–96. London: Oxford University Press.

Stuewig, J., and McCloskey, L. A. 2005. The relation of child maltreatment to shame and guilt among adolescents: Psychological routes to depression and delinquency. *Child Maltreatment*, *10*(4), 324–36.

Sunstein, C. R., Schkade, D., and Kahneman, D. 2000. Do people want optimal deterrence? *The Journal of Legal Studies*, *29*(1), 237–53.

Tabak, B. A., McCullough, M. E., Luna, L. R., Bono, G., and Berry, J. W. 2011. Conciliatory gestures facilitate forgiveness and feelings of friendship by making transgressors appear more agreeable. *Journal of Personality*, *80*(2), 503–36.

Tangney, J. P. 1995. Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, *38*(8), 1132–45.

Tangney, J. P., Miller, R. S., Flicker, L., and Barlow, D. H. 1996. Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, *70*(6), 1256–69.

Tangney, J. P., Stuewig, J., and Mashek, D. J. 2007. Moral emotions and moral behavior. *Annual Review of Psychology*, *58*(1), 345–72.

Tangney, J. P., Wagner, P., Fletcher, C., and Gramzow, R. 1992. Shamed into anger? The relation of shame and guilt to anger and self-reported aggression. *Journal of Personality and Social Psychology*, *62*4, 669.

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marschall, D. E., and Gramzow, R. 1996. Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. *Journal of Personality and Social Psychology*, *70*(4), 797–809.

Taylor, G. 1985. *Pride, Shame, and Guilt: Emotions of Self-Assessment*. Oxford: Clarendon Press.

Tetlock, P.E. 2002. Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological Review, 109*(3), 451-471.

Tittle, C. R. 1980. *Sanctions and Social Deviance: The Question of Deterrence*. New York: Praeger Publishers.

Tracy, J. L., and Robins, R. W. 2006. Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and Social Psychology Bulletin*, *32*(10), 1339–51.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., and Gee, J. O. 2002. Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes*, *89*(1), 839–65.

Ule, A., Schram, A., Riedl, A., and Cason, T. N. 2009. Indirect punishment and generosity toward strangers. *Science*, *326*(5960), 1701–4.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. 1987. Responsibility and the limits of evil: Variations on a Strawsonian theme. In F. D. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

Watson, G. 1996. Two faces of responsibility. *Philosophical Topics, 24*(2), 227-248.

Wenzel, M., Okimoto, T. G., Feather, N. T., and Platow, M. J. 2008. Retributive and restorative justice. *Law and Human Behavior*, *32*(5), 375–89.

Williams, B. 1985. *Ethics and the Limits of Philosophy*. London: Fontana Press.

Velleman, J. D. 2001. The genesis of shame. *Philosophy and Public Affairs*, *30*(1), 27–52.

Yang, M.-L., Yang, C.-C., and Chiou, W.-B. 2010. When guilt leads to other orientation and shame leads to egocentric self-focus: Effects of differential priming of negative affects on perspective taking. *Social Behavior and Personality: An International Journal*, *38*(5), 605–14.

Yzerbyt, V., Dumont, M., Wigboldus, D., and Gordijn, E. 2003. I feel for us: The impact of categorization and identification on emotions and action tendencies. *British Journal of Social Psychology*, *42*(4), 533–49.

Zebel, S., Zimmermann, A., Tendayi Viki, G., and Doosje, B. 2008. Dehumanization and guilt as distinct but related predictors of support for reparation policies. *Political Psychology*, *29*(2), 193–219. <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9221.2008.00623.x/full>

Zechmeister, J. S., Garcia, S., Romero, C., and Vas, S. N. 2004. Don't apologize unless you mean it: A laboratory investigation of forgiveness and retaliation. *Journal of Social and Clinical Psychology*, *23*(4), 532–64.

Zimbardo, P. G. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold and D. Levine (eds.), *1969 Nebraska Symposium on Motivation*, pp. 237–307. Lincoln, NE: University of Nebraska Press.