Ethic Theory Moral Prac (2009) 12:9–24
DOI 10.1007/s10677-008-9145-3

The Psychology of Dilemmas and the Philosophy of Morality

Fiery Cushman · Liane Young

Accepted: 3 December 2008 / Published online: 10 January 2009

Abstract We review several instances where cognitive research has identified distinct psychological mechanisms for moral judgment that yield conflicting answers to moral dilemmas. In each of these cases, the conflict between psychological mechanisms is paralleled by prominent philosophical debates between different moral theories. A parsimonious account of this data is that key claims supporting different moral theories ultimately derive from the psychological mechanisms that give rise to moral judgments. If this view is correct, it has some important implications for the practice of philosophy. We suggest several ways that moral philosophy and practical reasoning can proceed in the face of discordant theories grounded in diverse psychological mechanisms.

Keywords Moral psychology · Dilemmas · Trolley problem · Moral luck · Free will

1 Introduction: An Outsider's Perspective of Moral Philosophy

Our aim in this essay is to explore how current research in moral psychology has relevance to the work of moral philosophers. Being psychologists ourselves, we will begin with a brief sketch of our own outsiders' perspective on the landscape of moral philosophy. Two features stand out prominently. First, there are a number of basic kinds of moral theory that have persisted in a recognizable form for many generations—theories like virtue ethics, deontology, contractualism and utilitarianism. Specific details of each theory are malleable, but certain core concepts reliably attract philosophical attention. Second, there are a number of basic fault lines between moral theories that have also persisted in a recognizable form for many generations. These are often captured by moral dilemmas. Should the rights of

F. Cushman (✉)
Department of Psychology, Harvard University, 1484 William James Hall, 33 Kirkland St., Cambridge, MA 02138, USA
e-mail: cushman@wjh.harvard.edu

L. Young
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

one be sacrificed for the good of many? Can moral responsibility be reconciled with causal determinism? And so on.

Our thesis is that these two features of moral philosophy—divergent theories and persistent dilemmas—can be understood as the products of human psychology, and that such an understanding is of philosophical importance (see also Greene 2007; Sinnott-Armstrong 2007). Section 2 argues for this conclusion from a review of empirical moral psychology. Research suggests that a number of distinct psychological mechanisms accomplish moral judgment in ordinary people. These mechanisms sometimes conflict within a single individual, giving rise to the experience of a moral dilemma. Notably, it appears that ordinary people's mechanisms of moral judgment share core features with longstanding philosophical theories. A parsimonious account of these data is that key axiomatic claims grounding these philosophical theories are derived from standard psychological mechanisms, and therefore that philosophical moral theories conflict where standard psychological mechanisms conflict.

Section 3 explores the philosophical importance of this argument, focusing especially on cases of conflict between moral theories. We argue that this conflict is unavoidable—no moral theory can simultaneously satisfy the constraints of these multiple psychological systems. Nevertheless, we must still decide what to do when faced with moral dilemmas. Consequently, we recommend an expansive view of practical reasoning in which arguments grounded outside the moral domain help us adjudicate between moral demands in conflict. We also advocate the reframing or redesign of social institutions in order to avoid moral dilemmas in the first place.

2 A Multi-System Moral Psychology

If any single insight characterizes the current state of the field of moral psychology, it is that moral judgments are accomplished by multiple cognitive systems (Blair et al. 2006; Cushman et al. 2006; Greene et al. 2004, 2001; Haidt 2007; Koenigs et al. 2007; Pizarro and Bloom 2003; Young et al. 2007). This marks a notable shift in thinking. For its first 50 years, the field was dominated by theories positing a single system responsible for moral judgment. The work of decades of attribution theorists (e.g. Shaver 1985; Weiner 1995) took this perspective, focusing on the set of necessary and sufficient conditions for moral blame. These conditions were taken to be computed within a single unified system. Similarly, cognitive developmentalists, most notably Piaget (1965) and Kohlberg (1969), assumed the operation of a single system for moral judgment in their efforts to characterize patterns of developmental change.

The earliest challenges to these 'single process' views came from critics of Kohlberg, such as Gilligan (1982), who demonstrated consistent gender differences in moral reasoning, and Turiel (1983), who noted the coexistence of distinct mechanisms for reasoning about conventional violations and moral violations (i.e. "absolute" violations, by Turiel's definition of "moral"). More recently, attention has focused on the diverse roles of conscious reasoning versus unconscious intuition in moral judgment (Cushman et al. 2006; Haidt 2001; Pizarro et al. 2003) as well as the role of 'cold' cognition versus affect (Greene et al. 2004, 2001; Koenigs et al. 2007; Valdesolo and DeSteno 2006). A parallel movement is attempting to divide the moral domain according to the content of the judgment in question, postulating distinct mechanisms for judgments of help versus harm, for instance, or sexual taboo, or distributive justice (Blair et al. 2006; Haidt 2007; Schweder and Haidt 1993). Much work remains in evaluating the linkages between these theories, but it is clear

at least that the moral mind is a constellation of distinct cognitive process that can operate independently, often interact, and sometimes compete.

In this essay we focus on three cases that we think will be of particular interest to moral philosophers. In each case, the distinctions between cognitive systems suggested by empirical data parallel a prominent area of debate within the philosophical literature.

2.1 The Trolley Problem

A primary focal point of research in moral psychology is the trolley problem, introduced to modern philosophy by Foot (1967) and Thomson (1984). Numerous studies have demonstrated that a large majority of individuals consider it morally acceptable to use a switch to redirect a runaway trolley away from five victims and onto a single victim, but unacceptable to push a single victim in front of a runaway trolley in order to stop its progress towards five victims (Cushman et al. 2006; Greene et al. 2004, 2001; Koenigs et al. 2007; Mikhail 2000; Petrinovich et al. 1993; Valdesolo and DeSteno 2006). This pattern of judgments is consistent across a reasonably broad range of biological and cultural variation (Hauser et al. 2007).

Early work by Greene et al. (2001) demonstrated that relatively 'impersonal' scenarios, like the switch case, and relatively 'personal' scenarios, like the push case, yield dissociable patterns of neural activation. Specifically, impersonal scenarios characteristically yield greater activation in areas of the brain associated with effortful, deliberate reasoning, while personal scenarios yield greater activation in areas of the brain associated with emotion and social cognition. A follow-up study by Greene et al. (2004) demonstrated that the characteristic patterns of activation in impersonal cases can be used to predict subjects' responses to moral dilemmas that pose a choice between one life and many lives. For instance, subjects were asked whether it was appropriate for a mother to smother her crying baby in order to prevent the discovery of her hidden family by an enemy search party. Subjects who exhibited the greatest activity in areas associated with deliberate reasoning were more likely to endorse smothering the child.

Greene and colleagues have suggested that these characteristic patterns of brain activation reflect two distinct psychological processes of moral judgment: a cognitive system that favors welfare-maximizing choices, and an affective system that prohibits actions involving direct physical harm to specific individuals. Their diagnosis of the standard response to the trolley problem is that the switch case fails to activate the affective system, while the push case strongly activates it. Critically, in cases like the 'crying baby' dilemma, the two processes are thrown into conflict. Whichever process is more strongly activated determines the final moral judgment. Corroborating this dual-system account, patterns of brain activation in these cases reveal signatures of cognitive conflict: a neuronal reconciliation between the competing demands of separate psychological mechanisms.

Theories derived from neuroimaging data typically depend on correlations between brain activity and behavior. In order to test whether the brain regions implicated in 'personal' cases play a causal role in generating moral judgments, Koenigs, Young et al. (2007) investigated a population of individuals with damage to the ventromedial prefrontal cortex (vmPFC) who exhibited marked deficits in emotional processing of non-moral stimuli. Compared to healthy individuals and a control population of individuals with brain damage in other regions, the vmPFC individuals were significantly more likely to make welfare-maximizing decisions for moral dilemmas like the 'push' version of the trolley problem and the crying-baby case. On cases like the 'switch' version of the trolley problem, however, vmPFC individuals exhibited a perfectly normal pattern of responses. The implication of

these data is that the vmPFC contributes to the prohibition against direct harm that dominates in personal cases, but independent brain regions are responsible for moral judgments based on norms for welfare maximization that dominate in impersonal cases. Two additional studies of individuals with broadly similar neuropsychological profiles provide convergent evidence for this account (Ciaramelli et al. 2007; Mendez et al. 2005).

A natural interpretation of these two systems is that one produces patterns of judgments that conform to deontological rules (prohibiting direct harms against specific individuals) while the other engages explicit consequentialist reasoning (favoring welfare-maximizing choices). The divergent output of these two normative theories appears to correspond to distinct psychological systems in the minds of ordinary individuals. Moreover, this dual-system model of moral judgment provides a natural explanation for the phenomenological experience associated with a certain class of moral dilemmas exemplified by the 'crying baby' case. Dilemmas of this sort engage both systems, which elicit different judgments, resulting in cognitive conflict. The very debates carried out between individual philosophers who subscribe to one or another normative theory appear to be carried out between psychological systems, and within most ordinary individuals (Greene 2007).

2.2 Moral Luck

The question of whether a moral judgment should ever depend upon luck, famously posed by Williams (1981), has received considerable attention in the philosophical literature. There are several ways in which chance circumstances out of the agent's control might influence the moral standing of an agent or his or her action. We will focus on just one of these: a mismatch between the intentions behind an action and the consequences of that action. The dilemma posed by such cases is particularly stark in the case of negligent or reckless behavior. Nagel (1979) invites us to imagine two equally intoxicated people, each of whom decides to drive home. If a pedestrian walked out in front of either driver, a collision would be inevitable—just this occurs in the case of one driver, but the other driver makes it home without incident. On the one hand, it seems perverse to let the unlucky homicidal driver off with nothing more than a "driving under the influence" charge, or to punish the lucky driver as if he had killed. From this perspective, the drivers deserve different punishments. On the other hand, it may appear that two individuals who engaged in identical behavior do not deserve starkly different moral evaluations, solely on the basis of the unlucky timing of a pedestrian's stroll.

Indeed, recent studies implicate distinct psychological systems for moral judgments that focus either on the consequences of a behavior or the intentions underlying it. Evidence suggests that ordinary adults rely differentially on information about consequences and intentions when making judgments about moral wrongness versus punishment and blame (Cushman 2008). When evaluating whether an agent has acted wrongly, judgments are overwhelmingly dominated by information about the agent's intentions. Failed attempts at harmful behavior are judged as wrongful as completed attempts, while accidental harms are exonerated as if no harm occurred. But subjects' judgments of the blame and punishment deserved by those same agents are strikingly different: failed attempts are punished less harshly than completed attempts, and accidental harms are not fully exonerated, paralleling the law. Therefore, we may be inclined to blame and punish the two drunk drivers differently, on the basis of their lucky and unlucky outcomes, while, at the same time, evaluate the wrongness of their behavior similarly, on the basis of their similarly negligent and reckless attitudes and actions.

These data give us a fuller picture of the two processes of moral judgment that may be at play in cases of moral luck: a process that principally evaluates intentions and outputs

judgments of moral permissibility, and a process that is relatively more sensitive to consequences and outputs judgments of punishment. Notably, this view aligns with decades of research into the development of moral reasoning (Grueneich 1982; Hebble 1971; Kohlberg 1969; Piaget 1965; Shultz et al. 1986; Yuill and Perner 1988). Numerous studies have shown that young children have an early conception of morality centered on the concept of punishment, and are sensitive principally to information about the consequences of behavior. In the early elementary school years a shift occurs: children begin to understand morality in terms of duty, constraint, and reciprocity, and become more sensitive to information about the intentions underlying behavior. It appears that these two developmental stages may reflect an underlying cognitive architecture where distinct processes are at play. Indeed, preliminary data collected by Cushman and colleagues suggests that by age five children already rely more on outcomes when judging deserved punishment, and less when judging the naughtiness of a behavior.

Further evidence for a competitive interaction between two systems of moral judgment comes from a phenomenon termed 'blame-blocking' (Cushman 2008). Consider two people who attempt to murder a rival at a restaurant by sprinkling poppy seeds on his salad, believing that the rival to be allergic to poppy seeds. As it happens, both attempts fail: in each case, the rival is not allergic to poppy seeds at all, but instead to hazelnuts. In the "no harm" case, the rival goes on unaffected. But in the "harm" case, the rival happens to die by a totally causally independent mechanism: by eating hazelnuts placed in his salad by the unwitting chef. One might assume that the causally independent death of the rival would have no effect on how people punish the poppy-sprinkler, but in fact it does. People assign lesser punishment to the attempted murder by poppy seeds in the "harm" case compared to the "no harm" case. This result can be understood as the consequence of competition between one system of moral judgment that analyzes causal responsibility and another that analyzes mental culpability. When the chef causes death-by-hazelnuts, this absorbs causal blame for the crime, and people categorize the poppy-seed-sprinkler as 'off the hook' without considering his malicious intent. When no harm occurs, however, causal responsibility cannot be assessed. Consequently, the poppy-seed-sprinkler's punishment is fully determined by his malicious intent.

Although research into 'moral luck' and mismatches between intentions and consequences is still in its infancy, the emerging picture parallels the better-developed case of utilitarian vs. deontological moral judgment. There is evidence that distinct psychological processes are at play, that these can occasionally produce divergent outputs, and that moral judgments is sometimes a result of competition between these processes.

2.3 Moral Responsibility and Free Will

The problem of free will has been of increasing concern as scientific discovery uncovers increasing detail of the psychological and physical mechanisms underlying human behavior. The root of the problem is captured in a related family of questions. First, is determinism true? That is, are human decisions fully determined by past events? Second, given determinism, does free will exist? Third, given determinism, does moral responsibility exist? Our direct focus is on this third question, but in order to answer it we must consider the first two as well.

Among philosophers, incompatibilists claim that if human decisions are fully determined by past events, human agents do not have free will. Many philosophers extend incompatibilism to moral responsibility as in the third question above: human agents whose actions are fully determined by past events cannot be held morally responsible for

their actions. Compatibilists, by contrast, maintain that the truth of determinism does not undermine free will or moral responsibility; in other words, determinism and moral responsibility are not mutually exclusive (reviewed in Nichols and Knobe 2007).

Recent philosophical debate has turned to whether one view or the other is the natural or intuitive view, thus inviting an empirical contribution. Do folk intuitions consistently reflect either incompatibilism or compatibilism? Research into this question has greatly benefited from the pioneering efforts of a group of empirical philosophers (Nahmias et al. 2005; Nichols 2006; Nichols and Knobe 2007; Vargas 2005; Woolfolk et al. 2006). Early empirical work countered the standard position that the folk are incompatibilist (Kane 1999; Pereboom 2001), in revealing self-identified determinists to be just as punitive and retributivist as indeterminists (Viney et al. 1988, 1982). In other words, people who do harm ought to be punished and held morally responsible even if all of their actions are fully determined. This result suggests that belief in determinism allows for moral responsibility attributions typical of indeterminists. Consistent with this finding are the results of another study in which subjects read a rich description of a deterministic universe and then judged whether an agent in the universe was morally responsible for his misdeeds (Nahmias et al. 2005). For example:

Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00PM on January 26th, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195.

In response to this scenario, 83% of subjects judged Jeremy to be morally blameworthy for robbing the bank, which again speaks against the claim that folk intuitions are incompatibilist.

The question arises of how to resolve the standard position that folk intuitions are incompatibilist (e.g., if people's actions are determined, people are not morally responsible for their actions) and recent empirical data suggesting otherwise. A series of experiments by Nichols and colleagues (Nichols 2006; Nichols and Knobe 2007) proposes just the sort of multi-system solution that might be expected by our reader at this point in the essay. Nichols proposes that both philosophical parties may be right. That is, compatibilist intuitions may appear more frequently under some conditions and incompatibilist intuitions may appear under certain other conditions. According to this proposal (though there are certainly others), the key difference between these conditions is emotional salience of some sort; scenarios that provide enough concrete detail to elicit some level of emotional responding, similar to the one above, would tend to produce compatibilist intuitions.

Putting this proposal to the empirical test, Nichols and Knobe presented subjects with either affect-neutral or affect-laden descriptions of determinist universes and the subsequent questions respectively: "In Universe A, it is possible for a person to be fully morally responsible for their actions?" and "In Universe A, Bill stabs his wife and children so that he can be with his secretary. Is it possible that Bill is fully morally responsible for killing his family?" In both cases, subjects accepted the determinist terms of the example universe. However, the majority of subjects responded as incompatibilists to the first question posed in the affect-neutral context, and as compatibilists to the second question of the affect-laden

question. We leave aside the further question of whether the emotional outputs represent performance errors or moral competence and whether different descriptions of deterministic universes may result in subtly different patterns of judgments. We note only that such results showing that concrete emotional cases increase people's attributions of free will and moral responsibility (Nahmias et al. 2007) suggest the possibility that the problem of free will may reduce to the contribution of distinct psychological processes with distinct outputs. The rough sketch, which awaits further refinement, is that compatibilist intuitions may arise when emotional processes are engaged, while "cold" cognitive processes yield incompatibilist intuitions.

In keeping with the theme of this essay, the philosophical problem of this section, too, may be understood in the context of multiple underlying cognitive systems. The problem of free will and moral responsibility is thus reflected not just in distinct philosophical camps but potentially competing cognitive processes within individuals. Empirical research into this age-old philosophical topic is still coming into its own; however, a multi-system explanation of some sort may ultimately account for the phenomenological dilemma experience as well the centuries of philosophical discourse devoted to its resolution.

2.4 Conclusions

We have described three cases where moral judgments appear to be supported by multiple psychological systems. In each case, features of the psychological systems appear to parallel prominent positions in the philosophical literature, while the fault lines of cognitive conflict between psychological systems appear to parallel prominent philosophical debates. A parsimonious explanation of these parallels is that key axiomatic claims grounding philosophical moral theories are simply derived from the basic psychological mechanisms that accomplish moral judgment in ordinary people. Because there are multiple mechanisms, there are multiple theories; because the mechanisms sometimes conflict, the theories sometimes conflict.

3 Implications for Moral Theories

If this view is correct, we believe that it has some important implications concerning moral theories and the practice of philosophy. Thus, for the remainder of this essay we will explore what it would be like to conduct moral philosophy if the key axiomatic claims supporting moral theories are derived from psychological mechanisms of moral judgment.

3.1 Psychological Systems of Moral Judgment

The philosophical implications that we draw out below depend on a particular understanding of the standard psychological mechanisms that give rise to moral judgment, and the relations between these systems. These mechanisms, as we understand them, can be described as a set of axiomatic claims that can be applied to particular circumstances to yield judgments of value, duty, responsibility, retribution, fairness, and the like.[1] For

[1] Here, we do not attempt to define the scope of the moral domain, but instead rely on an intuitive sense of the sorts of judgments that have moral content. To paraphrase the Supreme Court's definition of pornography, "we know it when we see it".

instance, a 'deontological' mechanism of moral judgment could consist of the axiom, "It is prohibited to use harm to an individual as a means to an end." A distinct mechanism might yield judgments of moral responsibility, consisting of the axioms, "People are responsible only for actions under their control", and "People should be punished for wrongful acts for which they are responsible." Jointly, the output of these mechanisms might lead to the determination that Jane should be punished for intentionally killing her uncle for his inheritance. (The examples of axioms offered in this section are modeled on the research we presented above, but should be understood as simplified approximations.)

Our purpose in characterizing psychological mechanisms of moral judgment in terms of axiomatic claims is to make transparent the connection between these mechanisms and the formal moral theories that philosophers develop. This description probably provides a poor structural analogy for the computational and representational format of the underlying psychological processes. For example, characteristically deontological moral judgments may arise not by the application of an explicit rule prohibiting 'personal harms to specific individuals', but by the sensitivity of affective systems to salient, prototypical features of specific harmful actions. Nevertheless, these psychological mechanisms can be accurately translated into axiomatic terms, and doing so will help to clarify the relationship between psychological mechanisms and moral theories.

Importantly, different mechanisms of moral judgment can consist of axioms that yield opposing judgments. For instance, a consequentialist moral theory could consist of the axiom, "It is required to perform whichever action maximizes aggregate welfare". This consequentialist moral theory will occasionally demand behaviors that the deontological moral theory considered above would prohibit; smothering one's baby to prevent discovery by enemy soldiers is an excellent example. It should be clear that the consequentialist judgment in the 'crying baby' case is simply *not acceptable* under the deontological theory, and vice-versa. Each of these theories yields a definite answer to the 'crying baby' case, and those answers conflict.

One useful way to understand why these judgments are irreconcilable is to consider what a mechanism of moral judgment would have to look like if it *could* reconcile deontological and consequentialist concerns. Consider an alternative "joint" mechanism of moral judgment that reconciles deontological and consequentialist elements, and consisting of three axioms: "Subtract one point for every individual harmed as a means to an end", "Add 2 points for any action that maximizes aggregate welfare", and "Act so as to maximize points; toss a coin on ties". This mechanism yields a definite moral judgment, but not until it has adjudicated between its deontological and consequentialist elements. Because these elements are stated in terms of tradable 'points' rather than fixed moral demands, they do not conflict.

Yet, such adjudication is not possible when distinct deontological and consequentialist mechanisms make opposing moral *demands*. To be sure, we could posit a third mechanism that says, "If a behavior is prohibited on deontological grounds, subtract one point; if it is required on consequentialist grounds, add two points; act so as to maximize points, and toss a coin on ties". But the output of this third mechanism would still conflict with either the deontological or consequentialist judgment in the crying baby case! Those mechanisms (as we have provisionally defined them) do not contain any axiom deferring to the authority of a third, privileged mechanism. Nor do they output 'points' that merely encourage a particular conclusion. To the contrary, the output of each mechanism is a non-negotiable moral demand. Clearly a person possessed of two such demands must adjudicate between them, but the result of this adjudication will remain repugnant to one of the systems (at least).

To summarize our conclusions so far, conflict between mechanisms of moral judgment arises when several conditions are met. First, the mechanisms can each be characterized by distinct axioms. Second, the mechanisms produce opposing judgments in an identical category, such as a normative demand on action or a judgment of responsibility. Third, the demands on and judgments of behavior produced by each mechanism are inherently non-negotiable—these do not represent mere votes in favor of a particular conclusion but rather require a particular conclusion. Whether or not the particular psychological mechanisms we reviewed above meet these conditions remains an open, empirical question. We think there is substantial evidence that they do, and our arguments in the remainder of this section proceed on this assumption.

3.2 Conflict Within Individuals

In Section 2, we argued that conflict is a basic feature of the landscape of moral psychology. Specifically, we argued that moral dilemmas arise when distinct psychological mechanisms for moral judgment yield conflicting judgments of individual cases. Before turning to the matter of philosophical moral theories, we pause to consider the implications of our psychological account: intrapersonal conflict in the judgment of particular cases.

Within the lexicon of moral judgments—forbidden, permissible, obligatory, supererogatory and so forth—we suggest recognizing an additional type: the dilemma. Dilemmas may arise when moral questions that receive different answers from different psychological systems.[2] Note that a dilemma is not a moral question for which we have no answer, rather, it is a moral question for which we have multiple answers. Consider a mathematical analogy: the question "what is four divided by zero?" has no answer, whereas the question "what is the square root of four?" has two answers (2 and −2). Still better, consider the question "do the parallel lines A and B intersect?". This question also has multiple answers within a mathematical framework, and it depends on the axiomatic claims inherent to distinct mathematical systems: from a Euclidian perspective the answer is "no", and from a non-Euclidian perspective the answer is "yes". Moral dilemmas are cases where we arrive at differing answers to the question, "what is right?" by applying differing axiomatic claims.

Of course, acknowledging moral dilemmas as a basic type won't help anybody decide what particular course of action to pursue when confronted with one. One of the principal functions of a moral judgment is to guide practical reasoning, and to this end it will be necessary to choose between competing moral judgments. We shall have much more to say about this task below. Nevertheless, we think there are several reasons to remain conscious of dilemmas as a basic moral category, even if ultimately we must decide upon a single course of action. Doing so may direct our sympathy and support to those who face dilemmas, convince us not to reproach those whose chosen resolution differs from our own, and perhaps even afford us the opportunity to prevent dilemmas from arising in the first place. In any event, we suspect that the designation is worthwhile just insofar as it captures something familiar to everyday experience, is empirically supported by psychological research, and is abundantly represented in the philosophical literature.

[2] As noted by one of our reviewers, it may also be possible for a dilemma to arise when a single system yields contradictory demands. For instance, consider a system that prohibits any behavior that leads to the death of one's child. Now consider a dilemma in which one's children are held captive, and the captor explains that he will kill your son unless you tell him to kill your daughter. In this case, any behavior you perform stands in violation of the proposed psychological system. See also Tetlock's discussion of "tragic tradeoffs" (2003).

3.3 Conflict Between Theories

We have argued that philosophical moral theories are grounded in multiple psychological mechanisms that perform distinct computations and yield sometimes competing outputs. One way of putting this point is that philosophical moral theories formalize the axioms inherent to different mechanisms of moral judgment. If there were a single psychological computation responsible for moral judgment then the proper analysis of it might yield a single philosophical moral theory. Unfortunately, it seems not to work out that way. Consequently, those who fashion formal moral theories after human psychology face three options. First, they can develop 'pure' philosophical theories that formalize the axioms of a single psychological system. Second, they can develop 'indeterminate' philosophical theories that yield no definite answer to cases where psychological systems conflict. Third, they can develop 'hybrid' philosophical theories that attempt to trade off between the judgments of multiple psychological systems in order to reach a single, determinate answer. To put the point simply: if you have to two answers to a problem, you can accept one, neither, or both.

As a hypothetical case study in these three possibilities, let's suppose that the very simple characterizations of distinct deontological and consequentialist moral theories sketched above (Section 3.1) capture two basic psychological mechanisms that are widely shared. The first option available to philosophers is to develop 'pure' philosophical theories that formalize the axioms of either the deontological or the consequentialist psychological mechanisms. Deontological and consequentialist moral theories will inevitably conflict regarding particular cases. Nevertheless, the virtue of this approach is that the theories developed will fully satisfy half of our psychology.

The second option available to philosophers is to develop 'indeterminate' philosophical theories—theories that yield no judgment on cases where psychological systems conflict, but do yield judgments where psychological systems are congruent. An indeterminate theory would, therefore, yield no answer to the 'crying baby' case. However, a deontological psychological mechanism does not accept indeterminacy as a solution to the 'crying baby' case, rather, it *prohibits* smothering the baby. Likewise, a consequentialist psychological mechanism does not accept indeterminacy as a solution to the 'crying baby' case, rather, it *demands* smothering the baby. Thus, whereas 'pure' moral theories fully satisfy half our psychology in the crying baby case, 'indeterminate' moral theories satisfy neither. Moreover, certain cases of cognitive conflict may leave little room for congruent judgments. If determinism is true, compatibilism holds that we are morally responsible and incompatibilism holds that we are not. Unlike deontology and consequentialism, which are congruent on a large array of particular cases, compatibilist and incompatabilist judgments are in near-perfect conflict, rendering indeterminate theories a largely unattractive solution to the problem.

The third option available to philosophers is to develop 'hybrid' philosophical theories that trade off between the demands of distinct psychological systems. In a sense, reflective equilibrium is a process of hybrid theory construction: an attempt to balance and reconcile the conflicting demands of distinct psychological mechanisms. Hybrid moral theories have advantages. They can yield definite judgments in all cases, and the judgment yielded for any particular case is likely to converge with the judgment of at least one 'pure' psychological mechanism of moral judgment. But in cases of moral dilemmas, hybrid moral theories will still yield judgments that conflict with at least one 'pure' psychological mechanism of moral judgment.

Each of these options responds to conflict between competing psychological systems of moral judgment differently, but none of them can erase it. As we have seen, it is perfectly

feasible to construct a moral theory that yields a determinate judgment of any particular case. What is not possible is to construct a moral theory that can, for every case, simultaneously match the outputs of several psychological mechanisms of moral judgment. When you face a dilemma, no matter what you do, part of you is going to be dissatisfied.

3.4 What to do?

If we have the correct analysis of the role of moral psychology in the construction of philosophical moral theories, where does this leave the field of moral philosophy? One pessimistic view is that moral philosophers ought to become moral psychologists—or perhaps that they have just been moral psychologists all along. But something is clearly wrong with this view. Moral psychologists attempt to answer descriptive questions about how people decide what ought to be done, while moral philosophers attempt to answer normative questions about what actually ought to be done. True, we have argued that our minds do not provide a single satisfactory answer to the question, "what ought to be done?" for every case—that is, our multi-system moral psychology gives rise to genuine dilemmas. But in any case we must still do something! While descriptive psychological research has important implications for the task of practical reasoning (i.e. deciding what to do), it neither accomplishes nor obviates it. Above, we sketched three possible types of moral theory that could be constructed given the existence of multiple psychological mechanisms of moral judgment: 'pure' theories, 'indeterminate' theories, and 'hybrid' theories. What would moral philosophy look like—what would practical reasoning look like—under each of these approaches?

Philosophers who focus on developing pure moral theories may wish to avoid debates between moral theories and instead focus their efforts on progress within moral theories. By analogy, it is presumably more fruitful for folk singers and rappers to pursue perfection in their independent artistic pursuits than to debate which musical form achieves aesthetic superiority. And, insofar as pure moral theories are grounded in complex psychological mechanisms, there is plenty of room for distinct moral theories to be independently refined. Because this approach is tied closely to the operation of discrete psychological mechanisms, and because it abandons the claim of normative priority for any particular moral theory, it pushes the practice of moral philosophy firmly in the direction of a descriptive science. By the same token, it may provide an unconvincing basis for practical reasoning.

Setting aside indeterminate moral theories for the moment, let us turn to hybrid moral theories. The practice of trading off between competing moral demands is certainly familiar to philosophers—in essence, hybrid moral theories are the outcome of reflective equilibrium. Just as philosophers engage in reflective equilibrium to reconcile conflicting axioms in the development of a moral theory, there is good evidence that ordinary folk engage in a similar process of balancing conflicting moral demands in the moral judgment of particular cases. Research by Greene and colleagues shows that moral dilemmas are associated with longer periods of deliberation and the activation of brain regions associated with abstract reasoning and cognitive conflict (Greene et al. 2008, 2004, 2001). This provides some evidence for reflective equilibrium at the level of individual moral judgments.

We consider it very likely that ordinary folk engage in reflective equilibrium at the level of moral principles, as well. For instance, in one study we conducted subjects were asked to provide justifications for their prior moral judgments (Cushman et al. 2006). Sometimes we asked subjects to explain why they judged it less morally permissible to harm somebody by

physically contacting them than to harm them to an equivalent degree without physically contacting them. About 13%of subjects reversed their judgment when asked to provide a justification. For instance, one subject wrote, "I guess I was fooled by the line-pulling seeming more passive than the man-pushing, but that view is hard to justify now." In this case, the subject appears to be reflecting on the basis of her own moral intuition and rejecting it as a normatively valid consideration.

How is reflective equilibrium to proceed if we assume that it must decide between conflicting axiomatic claims in the absence of an adjudicating mechanism with moral authority? Let us begin by stating what reflective equilibrium *cannot* accomplish. Since it is supplied with conflicting axiomatic claims, it cannot hope to succeed by accepting all the axioms fully. Since the total set of moral axioms does not include internally consistent rules of priority, it cannot accept or reject some specific moral axioms by appeal to further moral axioms. Thus, in choosing among moral axioms by reflective equilibrium, we must look beyond the content of those axioms themselves. That is, we must look beyond psychological mechanisms that produce moral judgments, towards additional non-moral constraints on practical reasoning. For instance, we might choose among moral axioms by depending on prudential considerations (e.g. "when consequentialist axioms demand suicidal behavior, choose the deontological course of action"). Or, we might choose among moral axioms by depending on their experienced motivational force (e.g. "when retributive feelings exceed a threshold level, reject exculpation on grounds of causal determinism"). We might also choose among moral axioms by considering their psychological origins (e.g. "reject moral axioms the content of which appears to have been a target of natural selection").

This approach to moral philosophy engages in reflective equilibrium by using non-moral processes of practical reasoning to adjudicate between the conflicting demands of moral axioms. It surely demands a great deal of careful thought and argumentation. Moreover, it need not reduce to a descriptive project in moral psychology. Finally, in many ways, it seems to capture many core practices of moral philosophy as it currently exists. The distinct contribution of psychological research to this particular approach to moral philosophy is to make apparent (1) the origins of moral axioms (psychological mechanisms of moral judgment), (2) the impossibility of resolving conflict among these moral axioms, and (3) the necessary role of non-moral constraints on practical reasoning. The product of reflective equilibrium, on this view, is a general strategy for practical reasoning drawing from both moral and non-moral sources. As such, it might best be characterized not as making absolute moral demands, but rather as furnishing arguments in favor of particular choices.

Finally, we turn to indeterminate moral theories. At first glance, this approach appears deeply unsatisfying. In the case of moral dilemmas, indeterminate moral theories fail to provide any guidance to practical reasoning at all. Neither are indeterminate moral theories helpful in terms of developing a descriptively accurate characterization of any particular psychological mechanism of moral judgment. So, what would attract anybody to such an approach in the first place? In fact, indeterminate moral theories capture an important intuition: it is undesirable for a moral theory to demand behaviors that are categorically rejected by a psychological mechanism of moral judgment. It can be psychologically distressing for individuals, and it can lead to interpersonal conflict as well.

Can we derive useful practical guidance from this observation? If we take the psychological and interpersonal toll of moral dilemmas seriously, an important project that moral philosophy could adopt is to reduce their rate of incidence. This is essentially a project in social engineering, and we see two ways for it to proceed. One way is to change our psychological response to morally relevant situations, and the other way is to change

the situations themselves. (Note that these are not projects of developing indeterminate moral theories, they are simply projects motivated by the core impulse to take the indeterminate route and avoid the contentious demands of moral dilemmas in the first place).

The most direct approach to changing our psychological response to morally relevant situations is to alter our mechanisms of moral judgment at the level of their basic axiomatic claims. Moral philosophers could focus on developing arguments that reliably alter the psychological mechanisms of moral judgment, specifically attempting to make the output of diverse mechanisms compatible. To assess this possibility, we might first ask whether moral axioms ever change and second ask whether such change can be directed by education. Research in the cognitive development tradition, exemplified by Piaget and Kohlberg's theories of moral development, suggest that children's moral theories do undergo change at a fundamental level. For instance, both Piaget and Kohlberg argued that young children locate the source of normative value in authority, while older children locate the source of normative value in conventional agreements between individuals. However, we suspect that programs of moral education aimed at eliminating discord between distinct psychological systems will have severe practical limitations. Research by Haidt and colleagues (reviewed in Haidt 2001) suggests that some moral intuitions are highly resistant to counterargument (e.g. "sex between siblings is wrong"). In fact, the mere attempt to argue against so-called "sacred" or "protected" moral values can elicit outrage (Baron and Spranca 1997; Tetlock 2003). Certain moral commitments appear to be unshakable even while the explicit theories used to justify those commitments undergo substantial change.

A second way to avoid moral dilemmas is to change the way that morally relevant situations are psychologically framed. In fact, Haidt (2001) has proposed that successful moral argumentation often entails re-framing the input to mechanisms of moral judgment, rather than altering the axiomatic basis of the mechanisms themselves. For instance, in considering the moral status of abortion, we could frame the situation in terms of the rights of the mother or in terms of the rights of the fetus. In the case of some moral dilemmas, it may be possible to construct a frame in which the conflict between multiple systems of moral judgment disappears (see, e.g., Bartels and Medin 2007).

Finally, we can avoid moral dilemmas by changing the very structure of our social world. That is, we can choose to construct social institutions that avoid the conflict between multiple psychological systems of moral judgment. To illustrate how this might be accomplished, consider a very different sort of problem: designing social institutions that reduce the conflict between individuals over the exploitation of public goods such as the environment.

Some apartment buildings provide "free" utilities—gas, electricity and water. Of course, the utilities aren't really free at all. Landlords simply distribute the total cost of utilities evenly across all residents. The problem with this social institution is that it puts individuals' interests in conflict by providing a very small incentive for conservation. In a building with 100 apartments, every $1.00 of energy that an individual burns costs that individual only 1 cent. But at the same time, each resident pays the cost of all her neighbors' wasteful practices. Under this institution, residents consume a lot of energy, their rent is consequently very high, and the environment suffers. Nobody likes how anybody else behaves, but nobody has a financial incentive to change his or her own behavior. An alternative social institution bills residents individually for their utilities. Now residents pay $1.00 for every $1.00 they use. Under this institution residents consume less energy, their new bill for rent and utilities is consequently less than their former rent bill alone (on average), and the environment benefits as well. To put it another way, personal interests and

community interests become aligned. The lesson taken from this example is simple: we can choose to construct social intuitions that reduce certain types of interpersonal conflict.

In the case described above, conflict occurs between different individuals. By analogy, it may be possible to construct social institutions that avoid conflict within individual minds: that is, between distinct psychological mechanisms of moral judgment. Suppose, for the purpose of illustration, that affirmative action in higher education is clearly favored on utilitarian grounds because of the social benefits of diversity. Suppose, however, that it violates certain deontological intuitions when we witness a particular individual (in this case, the member of a privileged group) personally suffer the loss of admission to college. One institutional design for affirmative action is to rank all applicants without regard to race, gender, class, etc., and then to demote individuals from overrepresented groups and promote individuals from underrepresented groups. According to our toy model of utilitarian and deontological mechanisms, this would result in a dilemma: the utilitarian mechanism would endorse the system, while the deontological system would reject the system. An alternative institutional design for affirmative action is to raise scholarship money for an additional number of extra admissions each year, and then to select candidates for these extra slots only from underrepresented groups. By avoiding the specific demotion of any particular candidate for admission, such an institutional design might not be rejected by the deontological system. Put generally, by carefully constructing our social institutions, it may be possible to avoid the occurrence of some moral dilemmas. The design of effective social institutions is another possible task for moral philosophy, and one that has engaged philosophers for millennia. We suspect that progress in the field of moral psychology has much to offer this philosophical project.

4 Conclusion

Like others (Greene 2007; Sinnott-Armstrong 2007), we suspect that these psychological mechanisms play a direct role in the construction of moral theories. Current psychological research indicates that moral judgment is accomplished by multiple systems that sometimes yield conflicting outputs. Consequently, we are likely to be stuck with multiple moral theories that sometimes yield conflicting judgments. As a psychological phenomenon, we should expect this conflict to play out not only between individuals, but also within the minds of individuals.

The combination of a multi-system moral psychology and psychologically-grounded moral philosophy can explain the observation we made at the outset of this essay: there are certain stable and reliably attractive philosophical positions that yield conflicting judgments concerning particular cases. We have suggested three possible responses to this state of affairs. First, perfecting individual theories rather than choosing between them. Second, relying on non-moral considerations to help select between moral demands in the process of practical reasoning. Third, reshaping our own psychology (for instance, by argument and education), reframing moral conflicts, and redesign our social institutions in order to reduce the frequency with which moral dilemmas arise.

Motivated by our understanding of the specific philosophical dilemmas we presented above, we are attracted to the view that many philosophical dilemmas derive in some manner from disjunctions between competing mental processes (see also Greene 2007; Sinnott-Armstrong 2007). For instance, in the matter of reference: if the plays attributed to Shakespeare were written by the Earl of Oxford, who is Shakespeare: the man born in Stratford-upon-Avon, or the Earl of Oxford? This dilemma might arise from competition

between a mechanism that fixes reference to physical entities and a mechanism that fixes reference to functional roles. Or, in the matter of causation: if Nixon wouldn't have been impeached if dinosaurs still roamed the earth, did the extinction of the dinosaurs cause Nixon's impeachment? This dilemma might arise from competition between a mechanism that judges causation by mechanical production and a mechanism that judges causation by counterfactual dependency. Dilemmas like these provide the space for philosophical inquiry by fostering debate and guiding the explorations of a curious intellect. Ultimately, we suspect that they reflect the richly contoured landscape of the human mind.

Acknowledgements We wish to thank Richard Joyce, Walter Sinnott-Armstrong, and several reviewers for their valuable comments on this essay.

References

Baron J, Spranca M (1997) Protected values. Org Behav Hum Decis Process 70(1):1–16

Bartels D, Medin D (2007) Are morally motivated decision makers insensitive to the consequences of their choices? Psychol Sci 18(1):24–28 doi:10.1111/j.1467-9280.2007.01843.x

Blair RJR, Marsh AA, Finger E, Blair K, Luo J (2006) Neuro-cognitive systems involved in morality. Philos Explorations 9(1):13–27 doi:10.1080/13869790500492359

Ciaramelli E, Muccioli M, Ladavas E, di Pellegrino G (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Soc Cogn Affect Neurosci 2:84–92 doi:10.1093/scan/nsm001

Cushman FA (2008) Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition

Cushman FA, Young L, Hauser MD (2006) The role of conscious reasoning and intuitions in moral judgment: testing three principles of harm. Psychol Sci 17(12):1082–1089 doi:10.1111/j.1467-9280.2006.01834.x

Foot P (1967) The problem of abortion and the doctrine of double effect. Oxf Rev 5:5–15

Gilligan C (1982) In a different voice: Psychological theory and women's development. Harvard University Press, Cambridge

Greene JD (2007) The secret joke of Kant's Soul. In: Sinnott-Armstrong W (ed) Moral psychology. vol. 3. MIT, Cambridge

Greene JD, Morelli SA, Lowenberg K, Nystrom LE, Cohen JD (2008) Cognitive load selectively interferes with utilitarian moral judgment. Cognition.

Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron 44:389–400 doi:10.1016/j.neuron.2004.09.027

Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108 doi:10.1126/science.1062872

Grueneich R (1982) The development of childrens integration rules for making moral judgments. Child Dev 53(4):887–894 doi:10.2307/1129125

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834 doi:10.1037/0033-295X.108.4.814

Haidt J (2007) The new synthesis in moral psychology. Science 316(5827):998–1002 doi:10.1126/science.1137651

Hauser MD, Cushman FA, Young L, Jin R, Mikhail JM (2007) A dissociation between moral judgment and justification. Mind Lang 22(1):1–21 doi:10.1111/j.1468-0017.2006.00297.x

Hebble PW (1971) Devleopment of elementary school childrens judgment of intent. Child Dev 42(4):583–588 doi:10.2307/1127804

Kane R (1999) Responsibility, luck, and chance: reflections on free will and indeterminism. J Philos 96:217–240 doi:10.2307/2564666

Koenigs M, Young L, Adolphs R, Tranel D, Cushman FA, Hauser MD et al (2007) Damage to the prefrontal cortex increases utilitarian moral judgments. Nature 446:908–911 doi:10.1038/nature05631

Kohlberg L (1969) Stage and sequence: The cognitive-developmental approach to socialization. In: Goslin DA (ed) Handbook of socialization theory and research. Academic, New York, pp 151–235

Mendez MF, Anderson E, Shapria JS (2005) An investigation of moral judgment in frontotemporal dementia. Cogn Behav Neurol 18(4):193–197 doi:10.1097/01.wnn.0000191292.17964.bb
Mikhail JM (2000) Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'. Unpublished PhD, Cornell University, Ithaca.
Nagel T (1979) Mortal questions. Cambridge University Press, Cambridge
Nahmias E, Coates J, Kvaran T (2007) Free will, moral responsibility, and mechanism: experiments on folk intuitions. Midwest Stud Philos 31:214–242 doi:10.1111/j.1475-4975.2007.00158.x
Nahmias E, Morris S, Nadelhoffer T, Turner J (2005) Surveying freedom: folk intuitions about free will and moral responsibility. Philos Psychol 18(5):561–584 doi:10.1080/09515080500264180
Nichols S (2006) Folk intuitions about free will. J Cogn Cult 6:57–86 doi:10.1163/156853706776931385
Nichols S, Knobe J (2007) Moral responsibility and determinism: the cognitive sciencen of folk intuitions. Nous 41:663–685 doi:10.1111/j.1468-0068.2007.00666.x
Pereboom D (2001) Living without free will. Cambridge University Press, Cambridge
Petrinovich L, O'Neill P, Jorgensen MJ (1993) An empirical study of moral intuitions: towards an evolutionary ethics. J Pers Soc Psychol 64(3):467–478 doi:10.1037/0022-3514.64.3.467
Piaget J (1965) The moral judgment of the child. Free, New York
Pizarro DA, Bloom P (2003) The intelligence of the moral intuitions: comment on Haidt (2001). Psychol Rev 110(1):193–196 discussion 197–198
Pizarro DA, Uhlmann E, Bloom P (2003) Causal deviance and the attribution of moral responsibility. J Exp Soc Psychol 39:653–660 doi:10.1016/S0022-1031(03)00041-6
Schweder D, Haidt J (1993) The future of moral psychology: truth, intuition, and the pluralist way. Psychol Sci 4:360–365 doi:10.1111/j.1467-9280.1993.tb00582.x
Shaver KG (1985) The attribution of blame: Causality, responsibility, and blameworthiness.
Shultz TR, Wright K, Schleifer M (1986) Assignment of moral responsibility and punishment. Child Dev 57 (1):177–184 doi:10.2307/1130649
Sinnott-Armstrong W (2007) Abstract + Concrete = Paradox
Tetlock P (2003) Thinking the unthinkable: sacred values and taboo cognitions. Trends Cogn Sci 7(7):320–324 doi:10.1016/S1364-6613(03)00135-9
Thomson JJ (1984) The trolley problem. Yale Law J 94:1395–1415
Turiel E (1983) The development of social knowledge: Morality and convention. Cambridge University Press, Cambridge
Valdesolo P, DeSteno D (2006) Manipulations of emotional context shape moral judgment. Psychol Sci 17 (6):476–477 doi:10.1111/j.1467-9280.2006.01731.x
Vargas M (2005) The revisionist's guide to moral responsibility. Philos Stud 125(3):399–429 doi:10.1007/s11098-005-7783-z
Viney W, Parker-Martin P, Dotten SDH (1988) Beliefs in free will and determinism and lack of relation to punishment rationale and magnitude. J Gen Psychol 115:15–23
Viney W, Waldman D, Barchilon J (1982) Attitudes towards punishment in relation to beliefs in free will and determinism. Hum Relat 35:939–949 doi:10.1177/001872678203501101
Weiner B (1995) Judgments of responsibility : A foundation for a theory of social conduct. Guilford, New York
Williams B (1981) Moral luck. Cambridge Univeristy Press, Cambridge
Woolfolk RL, Doris JM, Darley JM (2006) Identification, situational constraint, and social cognition: studies in the attribution of moral responsibility. Cognition 100(2):283–301 doi:10.1016/j.cognition.2005.05.002
Young L, Cushman FA, Hauser MD, Saxe R (2007) The neural basis of the interaction between theory of mind and moral judgment. Proc Natl Acad Sci USA 104(20):8235–8240 doi:10.1073/pnas.0701408104
Yuill N, Perner J (1988) Intentionality and knowledge in childrens' judgments of actors responsibility and recipients emotional reaction. Dev Psychol 24(3):358–365 doi:10.1037/0012-1649.24.3.358