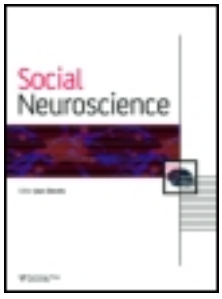

Social Neuroscience
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/psns20

Finding faults: How moral dilemmas illuminate cognitive structure

Fiery Cushman [a] & Joshua D. Greene [b]

[a] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA

[b] Department of Psychology, Harvard University, Cambridge, MA, USA

Available online: 23 Sep 2011

To cite this article: Fiery Cushman & Joshua D. Greene (2011): Finding faults: How moral dilemmas illuminate cognitive structure, Social Neuroscience, DOI:10.1080/17470919.2011.614000

To link to this article: http://dx.doi.org/10.1080/17470919.2011.614000

SOCIAL NEUROSCIENCE, 2011, i*First*, 1–11

Psychology Press
Taylor & Francis Group

Finding faults: How moral dilemmas illuminate cognitive structure

Fiery Cushman[1] and Joshua D. Greene[2]

[1]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA
[2]Department of Psychology, Harvard University, Cambridge, MA, USA

Philosophy is rife with intractable moral dilemmas. We propose that these debates often exist because competing psychological systems yield different answers to the same problem. Consequently, philosophical debate points to the natural fault lines between dissociable psychological mechanisms, and as such provides a useful guide for cognitive neuroscience. We present two case studies from recent research into moral judgment: dilemmas concerning whether to harm a person in order to save several others, and whether to punish individuals for harms caused accidentally. Finally, we analyze two features of mental conflict that apparently contribute to philosophical discord: the insistence that one answer to a problem must be correct ("non-negotiability") and the absence of an independent means of determining the correct answer ("non-adjudicability").

Keywords: Morality; Social neuroscience; Social cognition; Philosophy; Dilemmas; Trolley problem; Moral luck.

In philosophy a debate can live forever. Nowhere is this more evident than in ethics, a field that is fueled by apparently intractable dilemmas. To promote the well-being of many, may we sacrifice the rights of a few? If our actions are predetermined, can we be held responsible for them? Should people be judged on their intentions alone, or also by the consequences of their behavior? Is failing to prevent someone's death as blameworthy as actively causing it? For generations, questions like these have provoked passionate arguments and counterarguments, but few clear answers.

Here, we offer a psychological account of why philosophical dilemmas arise, why they resist resolution, and why scientists should pay attention to them. Building on a family of recent proposals (Cushman & Young, 2009; Greene, 2008; Sinnott-Armstrong, 2008), we argue that dilemmas result from conflict between dissociable psychological processes. When two such processes yield different answers to the same question, that question becomes a "dilemma." No matter which answer you choose, part of you walks away dissatisfied.

This explanation of philosophical dilemmas has an important payoff for psychological research, and we discuss two specific cases in which it has yielded promising results. In each case, social neuroscience has played an important role in distinguishing the psychological processes responsible for producing a dilemma. This, we suggest, is no accident; cognitive neuroscientific methods are particularly well suited to dissociating independent psychological processes (Henson, 2006). Consequently, philosophers' dilemmas provide a reliable guide to productive cognitive neuroscience by identifying the contours of distinct psychological process.

The research we review below focuses particularly on moral dilemmas, which is our own area of expertise. The psychological processes that contribute to moral

Correspondence should be addressed to: Fiery Cushman, CLPS Department, Box 1821, Brown University, Providence, RI 02912, USA. E-mail: fiery_cushman@brown.edu

Fiery Cushman thanks the Mind/Brain/Behavior Initiative for its generous support during the preparation of this work.

www.psypress.com/socialneuroscience http://dx.doi.org/10.1080/17470919.2011.614000

judgment are of interest in their own right, and play a central role in social cognition. But we conclude by pointing toward several areas where philosophical dilemmas appear to draw on core psychological competencies outside the moral domain. Our goal is to use case studies of moral judgment to illustrate a more general relationship between philosophy and psychology: Because philosophical debate erupts at the fault lines between psychological processes, it can reveal the hidden tectonics of the mind.

CASE 1: HARMING ONE TO SAVE MANY

One brand of vexing ethical dilemma arises when it is possible to deliberately harm an innocent person in order to promote the greater good of others. Consider a specific case. It is wartime. You and your fellow villagers are hiding from nearby enemy soldiers in a basement. Your baby starts to cry, and you cover your baby's mouth to block the sound. If you remove your hand, your baby will cry loudly, and the soldiers will hear. They will find you, your baby, and the others, and they will kill all of you. If you do not remove your hand, your baby will smother to death. Is it morally acceptable to smother your baby to death in order to save yourself and the other villagers?

Like many people, you may find it difficult to settle on an answer to this question. A growing body of research points to a particular explanation for this indecision: Two distinct processes of moral judgment provide contradictory answers (reviewed in Greene, 2008). One of these processes generates a strong, negative affective response to certain harmful actions; this process says, "Don't smother the baby!" The other process weighs the costs and benefits associated with an action in a controlled manner; this process says, "The baby will die no matter what; save yourself and the others." No matter which answer you settle on, part of your mind will reject it.

First, let's consider the emotional response to a certain class of harmful actions. Early evidence for this affective prohibition of harm relied on functional neuroimaging. Moral judgments made in response to dilemmas like the "crying baby" case, as compared to other cases in which the harm is less "personal,"[1] are associated with increased neural activity (as indexed by fMRI BOLD response) in regions associated with emotion, including subregions of the medial prefrontal cortex (mPFC) and the amygdala (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

Subsequent research demonstrates a causal role for the ventral (v) mPFC in prohibiting harmful actions. Specifically, individuals with damage to the vmPFC are far more likely than healthy individuals to endorse harmful behavior in order to promote a greater good (Koenigs et al., 2007; see also Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Mendez, Anderson, & Shapria, 2005). The precise role of the vmPFC relative to other brain regions is not fully established, however. Recent research suggests that the primary emotional responses to harmful action may originate in the amygdala, while the vmPFC's role may be to integrate this emotional response into an "all-things-considered" moral judgment (Shenhav & Greene, 2011). Moreover, medial regions of prefrontal cortex are implicated in a variety of other aspects of social cognition (Amodio & Frith, 2006), and so any reverse inference from patterns activation to underlying psychological mechanism must be approached cautiously.

As suggested above, certain harmful actions appear to engage this affective prohibition more than others. For instance, consider two cases in which a group of five persons' lives are threatened by an oncoming runaway trolley (Foot, 1967; Thomson, 1985). Most people will not endorse pushing somebody in front of the trolley in order to prevent it from hitting the people it now threatens. But, most people will endorse diverting the trolley onto a side-track where it will kill one person (Cushman, Young, & Hauser, 2006; Greene et al., 2001; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000). Two factors, in particular, seem to play a key role in distinguishing these cases (Greene et al., 2009; see also, Cushman et al., 2006). One is the physical nature of the harmful action. More specifically, when harm is caused by a direct transfer of a person's muscular force, this tends to evoke stronger moral condemnation. But, the factor of "personal force" only matters in conjunction with a second, more nuanced distinction (Greene et al., 2009; see also Cushman, et al., 2006). When harm is performed as a means to saving others ("This man could be used to stop the train"), it is judged morally worse than when harm is a side effect of saving others ("Diverting this train will also kill a man"). Across a variety of different situations, the joint presence of both factors—personal force, and means-to-an-end action—interacts to produce uniquely strong moral condemnation.

At first glance, this pattern of moral judgment is perplexing, but we suggest a possible explanation. Where in the mind would a representation like "the

[1] Greene et al. (2009) have revised their original (2001) definition of "personal." See also below.

application of direct muscular force as a means to an end" already exist, ready to adopt? Possibly, in a system that plans goal-directed actions to be effected by the voluntary flexion of skeletal muscles. Why might such motor "action plans"[2] constitute a basic input to the process of moral judgment? Perhaps if the process of moral judgment were designed to regulate one's own behavior, rejecting motor action plans that entail harm (Cushman, Gray, Gaffey, & Mendes, in press). Thus, the peculiar sensitivities of our affective response to others' harmful action may be best understood in light of a different purpose: to monitor our own behavior, sounding an emotional alarm when we plan to do harm.

According to this hypothesis, when we consider a mother faced with the task of smothering her own child, we imagine ourselves in her shoes. In doing so, we formulate the motor action plan that she must formulate: cupping our hand over the infant's mouth to silence him. We have a strong affective response against performing this action ourselves, and so we conclude that it would also be wrong for the mother to do this. This "simulated motor plan" hypothesis makes neuroscientific predictions; for instance, that the judgment of moral dilemmas like the crying baby case might be accompanied by increased activation in brain regions responsible for motor planning. It also entails that foreseen side effects (such as the harm that follows from redirecting the trolley) must not be part of the action plans in question and must therefore be represented elsewhere. Testing of these predictions is underway. But, whether or not the simulated motor plan account is validated, there is substantial evidence that some kind of negative affective response is triggered when we consider performing basic harmful actions.

Of course, if that were the end of the story, the "crying baby" case would not be a dilemma. Instead, smothering the baby would be judged unambiguously wrong. What makes the crying baby case a difficult dilemma is that a distinct process of moral judgment opposes the affective prohibition, instead endorsing the harmful act. Here, again, early evidence derived from functional neuroimaging. Greene and colleagues (2004) measured the neural activation evoked during moral judgment, comparing trials in which people endorsed harmful actions like smothering the baby (in order to bring about a greater good) to trials in which such harmful actions were condemned. They found that willingness to harm was associated with greater activation in regions associated with, among other things, cognitive control (Miller & Cohen, 2001) and thinking guided by explicit rules (Bunge & Wallis, 2007). These regions included the dorsolateral prefrontal cortex (DLPFC) as well as corresponding regions in the inferior parietal lobe. Thus, individuals who favor harming one person in order to save many others appear to be suppressing affective processes in favor of more effortful, controlled, rule-guided processes. The rule in question appears to be a utilitarian ("cost-benefit") one: A lesser harm is justified in the service of a greater good.

If controlled, rule-guided cognition is responsible for promoting the utilitarian conclusion that "the ends justify the means," then impairing one's capacity for controlled cognition should make one less utilitarian. Greene and colleagues (2008) tested this hypothesis by having subjects respond to cases like the "crying baby" dilemma while performing a cognitively demanding secondary task. Although subjects under cognitive load did not produce more utilitarian judgments than a control group, they did take significantly longer to make utilitarian judgments. By contrast, subjects' non-utilitarian judgments—hypothesized to rely on automatic emotional processing—were equally fast in both groups. Thus, the impairment of controlled cognition produced a selective effect on utilitarian moral judgment.

The same hypothesis makes parallel predictions about individual differences in cognitive style and their effects on moral judgment. That is, individuals who tend to rely more on controlled cognition and less on intuition (affective or otherwise) should give more utilitarian answers. As predicted, Bartels (2008) found that individuals with a more "rational" and less "intuitive" thinking style made more frequent utilitarian judgments, while Hardman (2008) found that individuals who scored high on the "cognitive reflection test" (Frederick, 2005) were about twice as likely to endorse using someone as a trolley stopper in the "footbridge" case and smothering the baby in the "crying baby" case.

By the same token, individuals with a blunted affective response to harmful or transgressive behavior should express greater willingness to smother the baby. Indeed, among a group of undergraduate students, those who indicated greater approval of utilitarian choices in cases like the "crying baby" also scored significantly higher on psychopathic traits and Machiavellianism (Bartels & Pizarro, 2011).

In summary, the data suggest that cases like that of the "crying baby" engage two distinct moral judgment processes. One produces a strong affective response

[2] The idea that "action plans" may play a key role in understanding moral judgment, and the trolley dilemma in particular, has been advanced by Mikhail (2007, 2000). Our proposal involves modifying Mikhail's theory in several ways, integrating it within a dual-process framework (Greene, forthcoming).

prohibiting harmful actions, perhaps by simulating and responding to the relevant "action plan." The other appears to rely on the controlled application of a utilitarian decision rule. When these systems conflict, mechanisms of cognitive control are engaged. The more forcefully individuals engage in controlled suppression of the affective prohibition, the more likely they are to endorse the utilitarian response. Critically, whichever answer is endorsed, one of the two systems is dissatisfied.

Can the data we have presented be explained by an alternative model in which there is no competition between distinct psychological processes? Moll, Oliveira-Souza, and Zahn (2008) propose one such alternative. They emphasize that an aversive reaction to direct harm must involve cognitive appraisal of the situation along with any affective response. Meanwhile, the utilitarian weighing of costs and benefits must entail some motivational force behind its brute logic. We agree (Cushman, Young, & Greene, 2010): Both resolutions to dilemmas like the crying baby case require elements of information processing (cognition) and motivation (affect).

Based on this observation, Moll and colleagues conclude that it is an *identical* combination of cognitive and affective mechanisms that is responsible for either resolution to dilemmas like the crying baby case (Moll et al., 2008). That is, the same cognitive and affective processes support our impulse not to smother the baby and also our impulse to save as many family members as we can. The dilemma strikes us as difficult simply because these two preferences cannot be simultaneously satisfied. The dual process model we have described differs on this point: It posits an essential distinction between automatic cognitive processes that trigger a strong affective response to direct harm, contrasting with controlled cognitive processes that suppress this affective response.

We have mentioned several of the factors that motivate this dual process account. It explains why neurological abnormalities can favor one particular resolution to a dilemma, rather than fostering general uncertainty or confusion (Ciarmarelli et al., 2007; Koenigs et al., 2007; Mendez et al., 2005). Likewise, it explains why cognitive load (Greene et al., 2008), individual thinking styles (Bartels, 2008), and psychopathic personality traits (Bartels & Pizarro, in press) are each associated with just one of the responses to the dilemma. It explains the distinct neural signatures associated with "personal" versus utilitarian considerations (Greene et al., 2001, 2004). Finally, it draws on the undeniably useful psychological distinction between automatic and controlled processes of decision-making and choice (Dayan & Niv 2008; Shiffrin & Schneider, 1977; Sloman, 1996). Of course, both kinds of process—automatic and controlled—must integrate affect and cognition. Moreover, there can be no doubt that automatic and controlled processes themselves are integrated in important ways. But the weight of the evidence suggests that they play dissociable roles in moral judgment and thereby contribute to the feeling of a dilemma.

Each of the psychological processes we have identified is mirrored in the philosophical literature. Deontological moral theories emphasize absolute, inviolable prohibitions against certain actions, where the prohibition is based on the nature of the action itself rather than its consequences. Kant (1785/1959), for example, regards the prohibition against using a person as a means to an end as central to morality. If we are correct, many details of deontological moral theories such as Kant's ultimately derive from features of the affective prohibition described above (Cushman, 2008b; Greene, 2008). Meanwhile, consequentialist moral theories such as the various forms of utilitarianism (Mill, 1863/1998) transparently formalize the process of explicit cost-benefit analysis, privileging it over other processes that yield contradictory answers. From a psychological perspective, it is little surprise that deontological and utilitarian philosophers have engaged in generations of debate over the relative merits of their theories without a clear victory for either side. The human mind furnishes not one, but two answers to the questions they attempt to answer.

This is precisely why philosophical debate illuminates cognitive structure. Identifying an action that is uncontroversially wrong (e.g., turning a trolley onto someone for fun) or uncontrovertibly required (e.g., turning a trolley away from five people and onto an empty track), helps us chart the general structure of judgments in the moral domain, but offers little guide to the unique signature properties of distinct psychological processes as they operate within the decision-making process. In contrast, points of philosophical tension—and intrapersonal conflict—provide an opportunity for differentiation. Thus, empirical investigations of moral dilemmas, despite many limitations, hold the promise of carving the mind at its joints.

CASE 2: PUNISHING ACCIDENTS

A second family of moral dilemmas concerns how we treat accidents; specifically, cases in which a person causes harm that he did not intend. Consider the following example. Two friends, Hal and Peter, share

beers over a Sunday afternoon football game at a local bar, and then each drives home. Both fall asleep, lose control of the wheel, and run off the road. Hall runs into a tree, but neither he nor the tree suffers great harm. Peter runs into a girl playing in her lawn and kills her. Should the accidental difference in the victim—the tree versus the girl—make a difference in our moral assessments?

This is sometimes called the problem of "moral luck," and it has produced decades of debate in philosophy and law (Hall, 1947; Hart & Honore, 1959; McLaughlin, 1925; Nagel, 1979; Williams, 1981). To see why, consider the legal consequences. In our home state of Massachusetts, Hal could expect a large fine and suspension of his license for driving while intoxicated. But Peter would face a mandatory minimum of 2.5 years in prison—and up to 15—for killing the girl. On the one hand, it seems unfair that these two friends are punished so differently for engaging in absolutely identical behavior. On the other hand, it seems unjust to sentence Hal to years in prison for drunk driving, or to let Peter off with a fine for killing a girl. No matter how you attempt to resolve the case, part of your mind recoils.

Recent research suggests that this dilemma, too, arises from a competitive interaction between two dissociable processes of moral judgment (Cushman, 2008a; Young, Cushman, Hauser, & Saxe, 2007). One is triggered by the occurrence of a harmful act and condemns the causally responsible individual, roughly in proportion to the amount of harm. The other considers individuals' mental states, condemning individuals whose actions foreseeably lead to harm, and exculpating individuals who cause harm that they could not have foreseen. These systems produce contradictory outputs when people cause harm accidentally: The outcome-based system is triggered by the harm and yields a negative moral assessment, while the mental-state system opposes this assessment because no harm was foreseen.

One source of evidence for a competitive interaction between outcome-based and foresight-based systems comes from studies of moral development. Children exhibit consistency across development in the judgment of attempted harms, but exhibit developmental change in the judgment of accidental harms (Costanzo, Coie, Grumet, & Farnill, 1973; Cushman, Sheketoff, Wharton, & Carey, 2011 in preparation; Zelazo, Helwig, & Lau, 1996). Specifically, young children condemn accidental harm-doers, while older children and adults typically do not. Early theories of moral development proposed a general shift from outcome-based reasoning to mental state-based reasoning over development (Kohlberg, 1969; Piaget, 1932/1965), but the simplest version of this theory cannot explain the data. After all, both attempted harms and accidental harms involve a mismatch between mental state and outcome; an attempted harm has foresight of harm without a harmful outcome, while an accidental harm has a harmful outcome in the absence of foresight.

The two-process account offered above suggests a possible resolution to this developmental puzzle. If the outcome process is triggered only in cases where harm actually occurs, then conflict between the outcome process and the mental state process will occur for accidental harms (+ outcome / – intent), but not for attempted harms (– outcome / + intent). Because the mental-state process's response to attempted harms is uncontested, young and older children judge these cases equivalently. But the mental-state process's response to accidental harms must overcome the outcome process's condemnation, and this capacity is attained over the course of development.

Neuroscientific studies provide further evidence that accidental harms create an intracranial conflict. Functional neuroimaging of adults engaging in moral judgment reveals activation of brain regions implicated in cognitive conflict and control for the judgment of accidental harms compared with non-accidental harms (Young et al., 2007). And, in a brain region associated with the representation of others' mental states, the right temporoparietal junction (RTPJ), fMRI BOLD responses during moral judgment predict the extent to which accidental harms are exculpated (Young & Saxe, 2009). These results suggest that overriding the outcome-based condemnation of accidental harms requires a robust representation of the harm-doer's mental state combined with effortful cognitive control.

Still, it might be objected that a robust representation of others' mental states is required to exculpate accidental harm even on a single process account of moral judgment. Both outcome information and mental state information clearly must be represented in the service of moral judgment. The question is whether these representations are seamlessly integrated prior to the computation of a moral judgment (as on a single-process model), or whether they separately support distinct moral judgments, which may then conflict (as on a two-process model). The developmental and neuroscientific evidence of cognitive conflict discussed above casts some doubt on the former view, but there is further evidence against it.

This evidence comes from a quirky pattern in our responses to attempted harms (Cushman, 2008a). Consider the following two cases. In the "no-harm" case, a runner attempts to poison his rival by sprinkling poppy seeds on his rival's salad. He believes

that his rival has a fatal allergy to poppy seeds, but, in fact, the allergy is to hazelnuts. Thus, his rival remains unharmed. The "coincidental harm" case is nearly identical, except that the chef just happens to have served the runner's rival a hazelnut salad. Thus, completely coincidentally, the rival dies. In this case, about 40% of subjects said that the runner deserved no prison time at all for his behavior, letting him off for attempted murder. By contrast, in the original case in which no one is harmed, only half as many subjects said that the runner deserves no prison time. Thus, the occurrence of a coincidental harm blocks people from assigning punishment to an attempted murder.

This result has a natural "dual-process" explanation. When no harm occurs, the outcome-based process is silent, and the runner is condemned on the basis of his malicious intent alone. But when a coincidental harm occurs, the outcome-based process is engaged. Causal responsibility for the rival's death is assigned to the salad, or perhaps to the chef, but certainly not to the scheming runner. Thus, although the scheming runner has murder on his mind in both cases, the outcome process exerts a countervailing force toward exoneration in the coincidental harm case only. Consequently, people are more likely to let the runner off the hook. It is much harder to accommodate this finding on a single-process model, however. If having a bad intention and causing a bad outcome seamlessly add up to determine punishment, then the sums should be equal for "no harm" and "coincidental harm": In both cases, the runner has a bad intention, but causes no harm. In order to explain why coincidental harms actually reduce punishment, it is not enough to say that the evaluation of outcomes adds up alongside the evaluation of mental states—it is necessary for outcomes to competitively block condemnation on the basis of intent.

The evidence from this coincidental harm case comes from an assessment of "deserved punishment," and the use of that particular dependent measure is no accident. Evidence suggests that the outcome-based process of moral condemnation plays a uniquely strong role in judgments of punishment, while playing a much more minor role in judgments of wrongness (Cushman, 2008a). Thus, for instance, people are likely to say that the drunk drivers Hal and Peter acted equally wrongly, regardless of what they hit. But, on the basis of different outcomes, people tend to assign different amounts of punishment. The role of accidental outcomes in punitive behavior is not a quirk of hypothetical judgment. Borrowing from methods in experimental economics, Cushman and colleagues (2009) have demonstrated the importance of accidental outcomes in laboratory games with real money on the table. When subjects use monetary punishments to respond to actual accidental behaviors, outcomes play a substantial role—larger, even, than intention.

These data suggest that neuroscientific studies of punitive behavior may afford the best opportunity to dissociate between outcome-based and mental-state-based processes of moral judgment. A recent neuroimaging study of punitive judgment points toward a promising line of research (Buckholtz et al., 2008). In keeping with studies reviewed above, the temporoparietal junction was associated with the exculpation of harms based on mental states. Meanwhile, amygdala activity was related to the degree of punishment, while dorsolateral prefrontal activity was related to the degree of responsibility for the crime. Much remains to be learned about the independent contributions of these regions to punitive judgment, but this initial research shows hope for insights from cognitive neuroscience.

Collectively, much of the data described above helps us to rule out an alternative, "single-process" explanation of moral luck offered by Royzman and Kumar (2004). In essence, their model posits that accidental outcomes do not directly impact moral judgments, but rather that accidental outcomes lead us to revise mental state attributions, in turn affecting our moral judgments. Thus, for instance, when we read about the drunk driver who kills a girl we think, "Of course—that was a very foreseeable outcome, and the agent is responsible for undertaking such a risky action," whereas when we read about the drunk driver who hits a tree we are less likely to consider the foreseeable harm to persons. This account surely explains *part* of the phenomenon of moral luck, as several studies indicate (Alicke & Davis, 1988; Young, Nichols, & Saxe, 2010). But the data we reviewed suggest that it cannot be a *complete* account of moral luck, and that diverse processes of moral judgment are at play. In particular, it cannot explain why judgments of punishment and wrongness are differentially sensitive to moral luck (Cushman, 2008a), or why moral luck is observed even when mental state information is perfectly known (Cushman et al., 2009), and it provides no explanation for the influence of coincidental outcomes on punishment for an attempted harm (Cushman, 2008a).

Here, again, a persistent philosophical dilemma paved the way for psychological and neuroscientific research. And, again, the dilemma arises from distinct psychological mechanisms that give different answers to the same question—in this case, whether to punish accidental outcomes. Consequently, the competing philosophical positions that theorists have taken in response to this dilemma may be best understood as reflections of the competing psychological processes.

WHAT MAKES DILEMMAS INTRACTABLE?

We have argued that philosophical dilemmas arise when distinct psychological processes give contradictory answers to the same question (see also Cushman & Young, 2009; Greene, 2008; Sinnott-Armstrong, 2008). But is our claim too broad? The mind is replete with conflict between psychological systems, and yet most of these conflicts have not resulted in years of philosophical debate. Why do some competitive processes give rise to intractable dilemmas while others do not? Here, we propose two key ingredients: Dilemmas arise when competing cognitive systems yield non-negotiable answers to questions that are not independently adjudicable.

To illustrate what we mean by "non-negotiable" and "non-adjudicable," we offer two examples of cognitive conflicts that do not give rise to corresponding philosophical dilemmas. The first is the conflict between our intuitive, "impetus" theory of physical motion and the explicit, tutored theory of Newtonian mechanics. A large body of psychological research suggests that these two mental systems for understanding physical motion can exist alongside each other and produce divergent responses (McCloskey, 1983; McCloskey, Caramazza, & Green, 1981). Yet there is no persistent, intractable philosophical debate over the merits of the two theories, and for obvious reasons. Empirical evidence clearly adjudicates in favor of Newtonian mechanics as a better approximation of reality than impetus theory. When a person is presented with divergent predictions about physical motion based on an impetus theory and Newtonian mechanics, these divergent predictions can be easily tested by interacting with the represented domain; that is, by running an experiment. Since both psychological mechanisms make a concrete prediction about the same phenomenon, exploring the phenomenon itself will decide between the theories.

In contrast, the moral cases described above cannot be adjudicated in this manner. The conclusions that "smothering the baby is absolutely wrong" and "smothering the baby is the best thing to do" do not make divergent predictions about the world that can be tested by an experiment—at least, not any experiment we know of. In the sense that they make predictions at all, those predictions concern how we will feel about smothering the baby. To say, "smothering the baby is wrong" predicts that smothering the baby will feel wrong. And, to say "smothering the baby is the best thing to do" predicts that smothering the baby will feel like the best thing to do. The difficulty is that both predictions are verified because they do not refer to a single phenomenon (e.g., how a single psychological system will respond to the dilemma). Rather, the predictions refer to different phenomena (how two different psychological systems will respond to the dilemma).

Of course, the need for adjudication does not arise for many moral claims. For instance, consider the claim, "It is morally impermissible to torture innocent children for fun." This claim makes the prediction that torturing children will feel impermissible, and not torturing children will feel permissible. From the perspective of any normal psychological process of moral judgment, this prediction is "verified." Consequently, people tend to unequivocally reject the claim that it is morally permissible to torture innocent children. In cases like this, it will often be the case that there are still dissociable systems at work; but it is hard to tell for sure, or to map the cognitive structure of each system. However, when we are confronted with a dilemma, it is very likely that dissociable systems are at work. Thus, we can use the dilemma as an entry point for understanding the cognitive structure of those systems.

Can we abstract away general properties of psychological systems that make them adjudicable versus non-adjudicable? Part of what makes conflict between impetus theory and Newtonian mechanics adjudicable is that they represent and predict a single, unitary set of phenomena. By contrast, part of what makes conflict between different moral systems non-adjudicable is that they represent and predict our own motivational states, which may be internally inconsistent. That is, the ordinary way of assessing whether smothering the baby is the right action or wrong action is to query the very psychological systems that disagree in the first place. Although not all representational systems produce adjudicable outputs, and some motivational systems may be adjudicable, we suspect that non-adjudicable conflicts arise most often in motivational systems.

This brings us to our second example, which falls squarely in the domain of motivation: the conflict between the preference for an immediate reward and the preference for the maximum reward. For example, ordering a steak may yield immediate gratification at the cost of one's future health and appearance, while ordering a salad may improve one's long-term health and looks, but at the cost of frustrating one's most salient, immediate desires. Cognitive neuroscientific research suggests that the competing preferences elicited by such situations are generated by independent psychological mechanisms (McClure, Laibson, Loewenstein, & Cohen, 2004). But there is not a persistent, intractable philosophical debate over

the merits of these two preferences.[3] More broadly, there are many cases in which we feel the tug of competing preferences, but these divergent motivations do not produce philosophical dilemmas. Why not?

When we choose between the steak and the salad, there must be some kind of higher-order mental mechanism that accomplishes the task—one that takes both preferences as input and yields a single choice as output. Our hypothesis is that this mechanism represents the preference for the steak not as, "this steak must be eaten," but rather as, "the value of eating the stake is such-and-so"—likewise for the salad. Thus, the competing preferences are viewed as inherently negotiable. This stands in contrast to the moral commitments, which are characterized by non-negotiability (Baron & Spranca, 1997; Tetlock, 2003). Our emotional prohibition of harm appears to takes the imperative form, "don't smother the baby!", not the preferential form "smothering the baby has such-and-so negative value"—likewise for the cognitive evaluation that a particular course of action is welfare-maximizing. (See also Greene, 2008, on "alarm" vs. "currency" emotions.)

Non-negotiability can also arise outside the moral domain. In fact, most claims about the fact of the matter—that is, the output of representational systems—will be non-negotiable. For instance, the determinist's claim that all human behavior is causally determined cannot be negotiated against the libertarian's insistence on the freedom of the will. These propositions are regarded as non-negotiable because each makes a definite claim about the actual state of affairs, flatly contradicting the other. Compare this with the case of the steak or the salad: it can be true that you want the steak while also being true that you want the salad. Choosing one does not logically contradict your desire for the other—it simply overrides it. Here, again, we can make a helpful generalization about motivation versus representational systems. Although not all motivational systems produce negotiable outputs (e.g., morality) and not all representational systems produce non-negotiable outputs, we suspect that non-negotiable conflicts most often arise in representational systems.

In summary, we have proposed two features that transform certain psychological conflicts into intractable dilemmas. When representational systems conflict, the conflict can often be adjudicated by independent observation of some property of the world. When motivational systems conflict, this conflict can often be negotiated by weighing the preferences against each other. Intractable dilemmas arise when psychological systems produce outputs that are non-adjudicable because they cannot be tested by independent observation, and non-negotiable because their outputs are processed as absolute demands, rather than fungible preferences. Our proposed recipe for dilemmas is speculative, however, and putting it to empirical test is an important matter for further research.

[3] Or at least not much of one. But, some philosophers have argued about whether it makes sense to discount the value of future rewards/punishments simply because they are in the future (e.g., Parfit, 1984, pp. 158–195).

TODAY'S PHILOSOPHY, TOMORROW'S SCIENCE

Throughout this essay, we have emphasized that philosophical dilemmas point the way toward productive cognitive neuroscience. To put our proposal to the test, we sketch in this section just a few of the philosophical dilemmas that we predict will play a key role in cognitive neuroscience, and in psychological research more broadly.

Determinism and responsibility

A vast philosophical literature engages the question of whether humans can be morally responsible if their behavior is causally determined. According to the thesis of causal determinism, every human action is ultimately caused by prior, external, or random factors. If this thesis is true, can humans be held morally responsible for their behavior? Research suggests that people are attracted to different answers to this question in different situations (reviewed in Nichols & Knobe, 2007). When phrased in the abstract, people tend to deny that moral responsibility is compatible with casual determinism. But, when embedded in a concrete, emotionally engaging case, people tend to assign moral responsibility even in the face of causal determinism. This suggests that different psychological processes may give rise to judgments of moral responsibility. Dissociating these processes may provide a richer psychological picture of human theories of action, choice, and responsibility (see also Sinnott-Armstrong, 2008).

Dualism

Are human consciousness and choice produced entirely by a physical system, the brain? Many people are drawn toward dualism, the view that certain properties of the mind must have some non-physical

basis. Evidence suggests that this intuition depends on a cognitive division of labor in the brain, according to which the behavior of intentional agents is processed in a fundamentally different stream than the behavior of non-intentional physical entities (Bloom, 2004). This division may be particularly strict in the infant brain. For instance, infants do not exhibit characteristic signs of surprise when intentional agents violate basic laws of physics by walking through walls, but exhibit robust surprise when non-intentional physical entities do so. Thus, a philosophical dilemma arises when the adult brain comes to recognize people as both intentional agents and physical entities (Greene & Cohen, 2004; Shariff, Greene, & Schooler, 2011 submitted). By finding cases where these competing systems conflict, we may be able to dissociate the rival processing streams that infants and adults use to assess intentional vs. non-intentional entities.

Double prevention

The philosophical literature on causation poses many dilemmas. Here, we focus on one: double prevention (Hall, 2004). Suppose a military bomber is on a mission to destroy a factory. It is escorted by a friendly fighter jet. An enemy fighter approaches, and takes aim at the bomber. But, in the nick of time, the friendly fighter shoots down the enemy. The bomber proceeds to drop its bombs on the factory, which is destroyed. Did the friendly fighter cause the factory to be destroyed? From one perspective, yes: if the friendly fighter had not shot down the enemy fighter, the factory would still be standing. It would not be unreasonable, for example, to award the friendly fighter pilot a medal for his efforts. From another perspective, no: the fighter had absolutely no physical connection to the factory at all. This dilemma seems to reveal rival processes of assigning causal responsibility, and cases like it may help to develop an account of how different mental systems accomplish causal attribution.

CONCLUSION

Building on several recent proposals (Cushman & Young, 2009; Greene, 2008; Sinnott-Armstrong, 2008), we have argued that philosophical dilemmas often arise when two distinct psychological processes yield conflicting answers to a common representational or motivational problem. Consequently, we suggest that philosophical dilemmas offer an important guide for psychological research generally, and for cognitive neuroscience specifically. We have presented two cases where this approach has already born fruit, and three cases where we suspect it holds great promise. We have also offered a speculative account of why certain instances of mental conflict give rise to dilemmas, while others do not. When conflict cannot be resolved by determining which answer is correct (adjudication), and cannot be resolved by balancing the relative value of the two answers (negotiation), the result is an intractable dilemma.

We conclude by turning our thesis on its head and asking, where will psychological research—and particularly cognitive neuroscience—lead the field of philosophy? There is a perspective from which our argument seems to undermine the foundation of philosophical debate. If rival claims simply reflect rival psychological systems, isn't the whole debate a charade? From this perspective, philosophers are just psychologists who take their conclusions too seriously, mistaking the psychology of the matter for the fact of the matter.

This perspective is attractive, but also flawed. To see why, let's return to the dilemma of the crying baby. Suppose the mother asks a neuroscientist, "What should I do?" The neuroscientist answers, "Funny thing—you have two mental approaches to this problem, and no matter what you do, one of them will be dissatisfied." That is descriptively correct, but it certainly does not help the mother. She must do something, and philosophy undertakes the unenviable task of helping her decide. To be sure, knowing why we feel an impulse toward one or another solution from a psychological perspective could play a critical role in helping us decide whether to favor one impulse over another. This is a point that we have emphasized elsewhere (Cushman & Young, 2009; Greene, 2008; Greene & Cohen, 2004). But scientific facts alone will never suffice to decide moral questions. Even as psychology and neuroscience map the geography of the moral mind, it is philosophers, policymakers, and ordinary people who must chart their own course.

Original manuscript received 16 May 2011
Revised manuscript accepted 28 July 2011
First published online day/month/year/

REFERENCES

Alicke, M., & Davis, T. (1988). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology*, *25*, 362–377.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*, 268–277.

Baron, J., & Spranca, M. (1997). Protected values. *Virology*, *70*(1), 1–16.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*, 381–417.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*. *121*(1): 154–161.

Bloom, P. (2004). *Descartes' baby*. New York, NY: Basic Books.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–935.

Bunge, S. A., & Wallis, J. D. (Eds.). (2007). *Neuroscience of rule-guided behavior*. New York, NY: Oxford University Press.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84–92.

Costanzo, P., Coie, J., Grumet, J., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development*, *44*(1), 154–161.

Cushman, F. A. (2008a). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F. A. (2008b). The origins of moral principles. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Cushman, F. A., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PLOS One*, *4*(8), e6699. doi: 6610.1371/journal.pone.0006699

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. (in press). Simulating murder: The aversion to harmful action. *Emotion*.

Cushman, F. A., Sheketoff, R., Wharton, S., & Carey, S. (2011). Excusing accidents across development: Evidence for a two-process model (manuscript in preparation).

Cushman, F. A., & Young, L. (2009). The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice*, *12*(1), 9–24.

Cushman, F. A., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. Doris & the Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 44–72). New York, NY: Oxford University Press.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (vol. 3). Cambridge, MA: MIT Press 35–80.

Greene, J. D. (forthcoming). *The moral brain and how to use it*. New York, NY: Penguin Group.

Greene, J. D., & Cohen, J. D. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*, 1775–1785.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Hall, J. (1947). *General principles of criminal law*. Indianapolis, IN: Bobbs-Merrill Company.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Hardman, D. (2008) Moral dilemmas: Who makes utilitarian choices? Unpublished manuscript, London Metropolitan University.

Hart, H. L. A., & Honore, T. (1959). *Causation in the law*. Oxford,UK: Clarendon Press.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language*, *22*(1), 1–21.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences*, *10*(2), 64–69.

Kant, I. (1785/1959). *Foundations of the metaphysics of morals*. (L. W. Beck, Trans.). New York, NY: Macmillan.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. A., Hauser, M. D., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908–911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151–235). New York, NY: Academic Press.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130.

McCloskey, M., Caramazza, A., & Green, B. (1981). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*, 1139–1141.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503–507.

McLaughlin, J. A. (1925). Proximate cause. *Harvard Law Review*, *39*(2), 149–199.

Mendez, M. F., Anderson, E., & Shapria, J. S. (2005). An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*(4), 193–197.

Mikahil, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Mikhail, J. M. (2000). Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice.' Doctoral dissertation, Cornell University, Ithaca, NY.

Mill, J. S. (1863/1998). *Utilitarianism*. New York, NY: Oxford University Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Moll, J., Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition: Sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, *1124*, 161–180.

Nagel, T. (1979). *Mortal questions*. Cambridge, UK: Cambridge University Press.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*, 663–685.

Parfit, D. (1984). *Reasons and persons*. New York, NY: Oxford University Press.

Piaget, J. (1932/1965). *The moral judgment of the child*. New York, NY: Free Press.

Royzman, E., & Kumar, R. (2004). Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio*, *8*(3), 329–344.

Shariff, A. F., Greene, J. D., & Schooler, J. W. (2011). His brain made him do it: Encouraging a mechanistic worldview reduces punishment (manuscript submitted for publication).

Shenhav, A., & Greene, J. D. (2011). Utilitarian calculations, emotional assessments, and integrative moral judgments: Differentiating neural systems underlying moral decision-making (manuscript in preparation).

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing. II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*(2), 127–190.

Sinnott-Armstrong, W. (2008). Abstract + concrete = paradox. In S. Nichols & J. Knobe (Eds.), *Experimental philosophy* (pp. 209–230). New York, NY: Oxford University Press.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*, 320–324.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*(6), 1395–1415.

Williams, B. (1981). *Moral luck*. Cambridge, UK: Cambridge University Press.

Young, L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–8240.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology. 1*(3), 333.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *48*, 2215–2220.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, *67*(5), 2478–2492.