Our multi-system moral psychology: Towards a consensus view

Fiery Cushman, Liane Young, and Joshua D. Greene

Introduction

In the field of moral psychology, a number of theoretical proposals that were at one time regarded as unconnected at best—and, at worst, contradictory—are showing signs of reconciliation. At the core of this emerging consensus is a recognition that moral judgment is the product of interaction and competition between distinct psychological systems. Our goal is to describe these systems and to highlight directions for further inquiry.

Recent research in moral psychology has focused on two challenges to the long-dominant cognitive development paradigm conceived by Piaget and nurtured by Kohlberg (Kohlberg, 1969; Piaget, 1965/1932; Turiel, 1983, 2005). The first challenge claims that moral judgment takes the form of intuition, accomplished by rapid, automatic and unconscious psychological processes (Haidt, 2001; Hauser, 2006: 1; Mikhail, 2000; Schweder & Haidt, 1993; see also Damasio, 1994: 165-201;), contra the cognitive developmentalists' assumption that moral judgment is the product of conscious, effortful reasoning. A central motivation for this challenge comes from studies demonstrating people's inability to articulate a rational basis for many strongly held moral convictions (Bjorklund, Haidt, & Murphy, 2000; Cushman, Young, & Hauser, 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000). The second and related challenge claims that moral judgment is driven primarily by affective responses (Blair, 1995; Damasio, 1994; Greene & Haidt, 2002; Schweder & Haidt, 1993), contra the cognitive developmentalists' focus on the deliberate application of explicit moral theories and principles to particular cases. Evidence for the role of affect is largely neuroscientific (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Damasio, 1994; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007; Mendez, Anderson, & Shapria, 2005), but also includes behavioral studies of moral judgment using affective manipulations (Valdesolo & DeSteno, 2006; Schnall, Haidt, Clore, & Jordan, in press; Wheatley & Haidt, 2005).

The evidence that moral judgment is driven largely by intuitive emotional responses is strong, but it does not follow from this that emotional intuition is the whole story. Concerning the role of intuition, the research of Kohlberg and others indicates a truly astonishing regularity in the development of explicit moral theories and their application to particular dilemmas (Kohlberg, 1969). Recent studies show that while people cannot offer principled justifications for some of their moral judgments, they are quite able to do so for others (Cushman et al., 2006), and that people alter some moral judgments when asked to engage in conscious reasoning (Pizarro, Uhlmann, & Bloom, 2003) or presented with appropriate opportunities (Lombrozo, in press). Others studies implicate reasoning processes in moral judgment using brain imaging (Greene et al., 2004) and reaction time data (Greene et al., in press). Together, these studies seem to

capture an important and relatively common experience: deliberation about right and wrong, informed by an awareness of one's explicit moral commitments.

In fact, the claim that moral judgment depends on affective responses at all has been met with skepticism, for instance by champions of "universal moral grammar" (Huebner, Dwyer & Hauser, 2008; Hauser, 2006; Mikhail, 2000). They observe that moral judgment requires computations performed over representations of agents, intentions, and causal relationships in order to output judgments of "right" and "wrong". Information processing of this sort necessarily precedes an affective response, the argument goes, and thus it must be that moral judgment that leads to affect, not the other way around.

Reconciling these apparent alternatives—intuitive versus rational[1], affective versus cognitive[2]—has therefore become a focal point of research. In our view, the most successful attempts share a common insight: moral judgment is accomplished by multiple systems. Here, we pursue a dual-process approach in which moral judgment is the product of both intuitive and rational psychological processes, and it is the product of what are conventionally thought of as "affective" and "cognitive" mechanisms. As we will see, a dual-process model of moral judgment can explain features of the data that unitary models cannot: dissociations in clinical populations, cognitive conflict in healthy individuals, and so on.

The new challenge we face is to understand the specific features of distinct systems and the processes of integration and competition among them. In this essay, we begin by reviewing the evidence in favor of a division between a cognitive system and an affective system for moral judgment. Next, we argue that the cognitive system operates by 'controlled' psychological processes whereby explicit principles are consciously applied, while affective responses are generated by 'automatic' psychological processes that are not available to conscious reflection. Thus, we suggest, the cognitive/affective and conscious/intuitive divisions that have been made in the literature in fact pick out the same underlying structure within the moral mind. Finally, we consider a set of Humean hypotheses according to which moral principles deployed in conscious cognition have affective origins.

The present essay focuses exclusively on two psychological systems that shape moral judgments concerning physically harmful behavior. Psychological researchers have often noted, however, that the moral domain encompasses much more than reactions to and prohibitions against causing bodily harm (Darley & Shultz, 1990; Gilligan, 1982/1993: 19; Haidt, 2007; Schweder & Haidt, 1993). Other sub-domains of morality might include the fair allocation of resources, sexual deviance, altruism and care, respect for social hierarchy, and religious devotion, for instance. Several authors have suggested that independent psychological systems are responsible for judgments in several of these

---

[1] By contrasting intuitive versus "rational" processes of moral judgment we aim to capture the common social psychological distinction between automatic (rapid, effortless, involuntary) and controlled (slow, effortful, voluntary) processes. Our purpose in choosing the term "rational" is not to imply the normative optimality of a particular decision, but rather to imply the use of deliberative reasoning in reaching that decision.
[2] For more on the somewhat artificial, but undeniably useful distinction between affect and cognition, see Greene (2008).

2

domains, and we regard such conjectures as plausible. Our focus on two systems that are important for judgments concerning harm is not presented as a complete account of moral psychology. On the contrary, we hope that the dual-process model explored here will eventually be understood as part of a larger constellation of psychological systems that enable the human capacity for moral judgment.

1. A dual-process model of moral judgment

Moral dilemmas come in many flavors, and a perennial favorite in moral philosophy forces a choice between harming one person and letting many people die, as in the classic trolley dilemmas (Foot, 1967; Thomson, 1985). In a case that we will call the switch dilemma, a runaway trolley threatens to run over and kill five people. Is it morally permissible to flip a switch that will redirect the trolley away from five people and onto one person instead, thus saving five lives at the cost of one? In a large web survey, most people said that it is (Hauser et al., 2007). This case contrasts with the footbridge dilemma. Here, one person is standing next to a larger person on a footbridge spanning the tracks, in between the oncoming trolley and the five. In this case, the only way to save the five is to push the large person off of the footbridge and into the trolley's path, killing him, but preventing the trolley from killing the five. (You can't stop the trolley yourself because you're not big enough to do the job.) Most people surveyed said that in this case trading one life for five is not morally permissible. It appears that most ordinary people, like many philosophers, endorse a characteristically consequentialist judgment ("maximize the number of lives saved") in first case and a characteristically deontological judgment ("harm is wrong, no matter what the consequences are") in the second. (For a discussion of why we consider it legitimate to refer to these judgments as "characteristically deontological" and "characteristically consequentialist," see Greene (2007)). This pair of dilemmas gives rise to the "trolley problem", which, for decades, philosophers have attempted to solve (Fischer & Ravizza, 1992; Kamm, 1998, 2006). What makes people judge these cases differently?

Greene and colleagues' dual-process theory of moral judgment (Greene et al., 2001, 2004, 2007) attempted to characterize the respective roles of affect and controlled cognition in people's responses to these dilemmas (Greene et al, 2001). Specifically, Greene and colleagues proposed that the thought of harming someone in a "personal" way, as in the footbridge dilemma, triggers a negative emotional responses that effectively says, "That's wrong, don't do it!". According to their theory, this emotional alarm bell overrides any consequentialist inclination to approve of the five-for-one trade-off. In contrast, people tend to say that redirecting the trolley in the switch case is morally permissible because the "impersonal" nature of this action fails to trigger a comparable emotional response. In the absence of such a response, consequentialist moral reasoning ("Five lives are worth more than one") dominates the decision. The core insight of this proposal is that our discrepant judgments in the switch and push cases may not come from the operation of a single psychological system, but rather from competition between two distinct psychological systems.

Putting this proposal to the empirical test, Greene and colleagues examined the neural activity of people responding to various "personal" and "impersonal" moral dilemmas. As predicted, they found that brain regions associated with emotion (and

social cognition more broadly) exhibited increased activity in response to "personal" moral dilemmas such as the footbridge case. These brain regions included a region of the medial prefrontal cortex (Brodmann's area 9/10) that was damaged in the famous case of Phineas Gage, the Nineteenth Century railroad foreman whose moral behavior became severely disordered after a tragic accident sent a metal tamping iron through his eye socket and out the top of his head (Damasio, 1994: 3; Macmillan, 2000). In contrast, and also as predicted, Greene and colleagues found that brain regions associated with controlled cognitive processes such as working memory and abstract reasoning exhibited increased activity when people were responded to "impersonal" moral dilemmas such as the switch case.

Building on this finding, Greene and colleagues conducted a second study (Greene et al., 2004). Whereas their first study identified patterns of neural activity associated with the kind of dilemma in question ("personal" vs. "impersonal"), the second study identified patterns of neural activity associated with the particular judgments people made ("acceptable" vs. "unacceptable"). They focused their analysis on difficult dilemmas in which harming someone in a "personal" manner would lead to a greater good. Here is an example of a particularly difficult case, known as the crying baby dilemma:

> *"Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside, you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others, you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?"*

Subjects tend to take a long time to respond to this dilemma, and subjects' judgments tend to split fairly evenly between the characteristically consequentialist judgment ("Smother the baby to save the group") and the characteristically deontological judgment ("Don't smother the baby"). According to Greene and colleagues' dual-process theory, the characteristically deontological responses to such cases are driven by prepotent emotional responses that nearly everyone has. If that's correct, then people who deliver characteristically consequentialist judgments in response to such cases must override their emotional responses.

This theory makes two predictions about what we should see in people's brains as they respond to such dilemmas. First, we would expect to see increased activity in a part of the brain called the anterior cingulate cortex (ACC), which, in more dorsal subregions, reliably responds when two or more incompatible behavioral responses are simultaneously activated, i.e. under conditions of "response conflict" (Botvinick, Braver, Barch, Carter, & Cohen, 2001). For example, if one is presented with the word "red" written in green ink, and one's task is to name the color of the ink, then one is likely to experience response conflict because the more automatic word-reading response ("red") conflicts with the task-appropriate color-naming response ("green") (Stroop, 1935).

According to this dual-process theory, the aforementioned difficult dilemmas elicit an internal conflict between a prepotent emotional response that says "No!" and a consequentialist cost-benefit analysis that says "Yes." Consistent with this theory, Greene and colleagues found that difficult "personal" dilemmas like crying baby case elicit increased ACC activity, relative to easier "personal" dilemmas, such as whether to kill your boss because you and others don't like him, in which reaction times are shorter and judgments are more overwhelmingly negative. Second, we would expect to see increased activity in a part of the brain known as the dorsolateral prefrontal cortex (DLPFC). This part of the brain is the seat of "cognitive control" (Miller & Cohen, 2001) and is necessary for overriding impulses and for "executive function" more broadly. Once again, if the characteristically deontological judgment is based on an intuitive emotional response, then giving a characteristically consequentialist response requires overriding that response, a job for the DLPFC. As predicted, Greene and colleagues found that consequentialist judgments in response to difficult "personal" dilemmas ("Smother the baby in the name of the greater good") are associated with increased activity in the DLPFC relative to the activity associated with trials on which deontological judgments were made.

These neuroimaging results support a dual-process theory of moral judgment in which distinct "cognitive" and emotional processes sometimes compete. But a consistent association between neural activity and behavior cannot provide conclusive evidence that the neural activity is a cause of the behavior. To provide stronger evidence for such causal relationships, one must intervene on the neural process. In a recent study, Greene and colleagues (2008) did this by imposing a "cognitive load" on people responding to difficult "personal" moral dilemmas like the crying baby dilemma. People responded to the moral dilemmas while simultaneously monitoring a string of digits scrolling across the screen. The purpose of this manipulation is to disrupt the kind of controlled cognitive processes that are hypothesized to support consequentialist moral judgments. They found, as predicted, that imposing a cognitive load slowed down characteristically consequentialist judgments, but had no effect on characteristically deontological judgments. (Deontological judgments were in fact slightly faster under cognitive load, , but this effect was not statistically significant.)

A complementary tactic is to manipulate the relevant emotional responses, rather than the capacity for cognitive control. Valdesolo and DeSteno (2006) did this by presenting either comedic video clips or affectively neutral video clips to two groups of subjects who then responded to versions of the switch and footbridge dilemmas. They reasoned as follows: If people judge against pushing the man in front of the trolley because of a negative emotional response, than a dose of positive emotion induced by watching a short comedic sketch from *Saturday Night Live* might counteract that negative response and make people's judgments more consequentialist. As predicted, they found that people who watched the funny video were more willing to endorse pushing the man in front of the trolley.

Yet another method for establishing a causal relationship between emotion and moral judgment is to test individuals with selective deficits in emotional processing. This approach was first taken by Mendez and colleagues (2005) in a study of patients with frontotemporal dementia (FTD), a disease characterized by deterioration of prefrontal and anterior temporal brain areas. FTD patients exhibit blunted emotion and diminished

regard for others early in the disease course. Behavioral changes include moral transgressions such as stealing, physical assault, and unsolicited or inappropriate sexual advances. Mendez and colleagues presented FTD patients with versions of the switch and footbridge dilemmas and found, as predicted, that most FTD patients endorsed not only flipping the switch but also pushing the person in front of the trolley in order to save the five others. Mendez and colleagues suggest that this result is driven by the deterioration of emotional processing mediated by the ventromedial prefrontal cortex (VMPC). Since neurodegeneration in FTD affects multiple prefrontal and temporal areas, however, firm structure-function relationships cannot be ascertained from this study.

To fill this gap, moral judgment has since been investigated in patients with more focal VMPC lesions—that is, lesions involving less damage to other structures. Like FTD patients, VMPC lesion patients exhibit reduced affect and diminished empathy, but unlike FTD patients, VMPC lesion patients retain broader intellectual function. Thus, VMPC patients are especially well-suited to studying the role of emotion in moral judgment. Koenigs, Young, and colleagues (2007) tested a group of six patients with focal, adult-onset, bilateral lesions of VMPC to determine whether emotional processing subserved by VMPC is in fact necessary for deontological moral judgment. In this study patients evaluated a series of impersonal and personal moral scenarios, used by Greene and colleagues in the neuroimaging work discussed above. VMPC patients responded normally to the impersonal moral scenarios, but for the personal scenarios the VMPC patients were significantly more likely to endorse committing an emotionally aversive harm (e.g., smothering the baby) if a greater number of people would benefit. That is, they were more consequentialist. A second lesion study conducted by Ciaramelli and colleagues (2007) produced consistent results. Thus, these lesion studies lend strong support to the theory that characteristically deontological judgments are – in many people, at least – driven by intuitive emotional responses that depend on the VMPC, while characteristically consequentialist judgments are supported by controlled cognitive processes based in the DLPFC.

It is worth noting that this dual-process theory, and the body of evidence that supports it, run counter to the traditional philosophical stereotypes concerning which normative positions are most closely allied with emotion vis-à-vis controlled cognition. Historically, consequentialism is more closely allied with the "sentimentalist" tradition (Hume, 1739/1978), while deontology is more closely associated with Kantian (1785/1959) rationalism. According to the dual-process theory, this historical stereotype gets things mostly backwards. The evidence suggests that characteristically deontological judgments are driven primarily by automatic emotional responses, while characteristically consequentialist judgments depend more on controlled cognition (Greene, 2007). Of course, this association between emotion and deontology, on the one hand, and consequentialism and controlled cognition, on the other, is bound to be overly simple. In what follows, we present a more nuanced picture of the interactions between cognition and affect that support both characteristically consequentialist and characteristically deontological moral judgments.

2. Intuition and affect: a common system?

The division between affective and cognitive systems of moral judgment proposed by Greene and colleagues is by now well-supported. Several critics have noted, however, that early formulations of the theory left many details of the affective system unspecified (Cushman et al., 2006; Hauser, 2006: 8; Mikahil, 2007). Taking the specific example of the trolley problem, what is it about the footbridge dilemma that makes it elicit a stronger emotional responses than the switch dilemma? As Mikhail (2000) observed, this question can be answered in at least two complementary ways. On a first pass, one can provide a descriptive account of the features of a given moral dilemma that reliably produce judgments of "right" or "wrong". Then, at a deeper level, one can provide an explanatory account of these judgments with reference to the specific cognitive processes at work.

A natural starting point for the development of a descriptive account of our moral judgments is the philosophical literature, which in the last half-century has burgeoned with principled accounts of moral intuitions arising from hypothetical scenarios (e.g. Fischer & Ravizza, 1992). Among the most prominent accounts of the trolley problem is a moral principle called the "Doctrine of Double Effect", or DDE. According to proponents of the DDE, the critical difference between the switch and footbridge cases is that the large man in the footbridge case is *used* to stop the train from hitting the five, whereas the death of the victim in the switch dilemma is merely a *side-effect* of diverting the trolley away from the five. In its general form, the DDE states that it is impermissible to use a harm as the means to achieving a greater good, but permissible to cause a harm as a side-effect of achieving a greater good. Setting aside its validity as a moral principle, we may ask whether the DDE is an accurate descriptive account of the moral judgments of ordinary people.

The characteristic pattern of responses elicited by the trolley problem provides some evidence for this view, but the DDE is certainly not the only account for subjects' judgments of these cases. Another obvious distinction between the footbridge and switch cases is that the former involves physical contact with the victim, while the latter occurs through mechanical action at a distance (Cushman et al., 2006). At a more abstract level, the footbridge case requires an intervention on the victim (pushing the large man), while the bystander case merely requires an intervention on the threat (turning the trolley) (Waldman & Dieterich, 2007).

In order to isolate the DDE from these sorts of alternative accounts, Mikhail (2000) tested subjects on a modified version of the switch dilemma. Both cases involve a "looped" side track that splits away from, and then rejoins, the main track (Figure 1). Five people are threatened on the main track, and they are positioned beyond the point where the side track rejoins. Thus, if the train were to proceed unimpeded down either the main track or the side track, the five would be killed. In the analog to the footbridge case, there is a man standing on the side track who, if hit, would be sufficiently large to stop the train before it hits the five. In the analog to switch case, there is a weighted object on the side track which, if hit, would be sufficiently large to stop the train before it hits the five. However, standing in front of the weighted object is a man who would first be hit and killed. These cases preserve the distinction between harming as a *means* to saving five (when the man stops the train) and harming as a *side effect* of saving five (when an object stops the train and a man dies incidentally). However, neither involves physical contact, and both require a direct intervention on the trolley rather than the victim.

Mikhail found that subjects reliably judged the looped means case to be morally worse than the looped side effect case, although the size of the effect was markedly smaller than the original switch / footbridge contrast. These results suggest that the DDE is an adequate descriptive account for at least some part of the moral distinction between the fat man and bystander cases. A follow-up study involving several thousand subjects tested online replicated this effect, and found it to be remarkably consistent across variations in age, gender, educational level, exposure to moral philosophy and religion (Hauser et al., 2007). While the specific wording of Mikhail's "loop" cases has been criticized (Greene et al., in preparation; Waldman & Dieterich, 2007), subsequent research has demonstrated use of the DDE across multiple controlled pairs of moral dilemmas (Cushman et al., 2006).

These initial studies establish that DDE is at least part of an accurate description of subjects' moral judgments, but leave open the explanatory question: how is the means/side-effect distinction manifested in the psychological systems that give rise to moral judgment? Several related proposals have been offered (Cushman, Young, & Hauser, in preparation; Greene et al., in preparation; Mikahil, 2007), but we will not discuss these in detail. All three share in common the claim that when harm is used as the means to an end it is represented as being intentionally inflicted to a greater degree than when harm is produced as a side-effect. Cushman and colleagues (in preparation) provide evidence that that this relationship holds even in non-moral cases. That is, when somebody brings about a consequence as a means to accomplishing a goal, participants were more likely to judge that the effect was brought about intentionally than when somebody brings about a consequence as a foreseen side-effect of their goal. Many details remain to be worked out in providing a mechanistic account of the DDE in producing moral judgments, but it is likely to involve representations of others' mental states.

Several studies have explored another dimension of the cognitive processes underlying the DDE: specifically, whether they operate at the level of conscious awareness. Hauser et al (2007) selected a subset of participants who judged killing the one to be impermissible in the loop means case but permissible in the loop side-effect case and asked them to provide a justification for their pattern of judgments. Of twenty-three justifications analyzed, only three contained a principled distinction between the two cases (e.g. "… the man's death is a byproduct rather than direct goal of trying to save the five men"). Subsequent research has replicated this result with a much larger group of participants, demonstrating the general inability of a majority of individuals to provide a sufficient justification for a variety of double effect cases (Cushman et al., 2006).

These results illustrate a phenomenon that Haidt has termed "moral dumbfounding" (Bjorklund et al., 2000). Moral dumbfounding occurs when individuals make moral judgments that they confidently regard as correct, but then cannot provide a general moral principle that accounts for their specific judgment. One example is the moral prohibition against incest among consenting adults. Haidt and colleagues (2000) told subjects about a brother and sister who make love, but who use highly reliable contraception to avoid pregnancy, are rightly confident that they will suffer no negative emotional consequences and they will be able to keep their activities private, etc. Presented with such cases, subjects often insist that the siblings' actions are morally

wrong, despite the fact that they cannot provide more general justification for this judgment.

Haidt and colleagues argue that these difficult-to-justify moral judgments are generated by rapid, automatic, unconscious mental processes—in short, intuitions. Moreover, research indicates that these intuitions are supported by affective processing. For instance, subjects' moral judgments are harsher when made in a physically dirty space (Schnall, Haidt, Clore, & Jordan, in press). It appears that subjects' feelings of disgust brought about by the dirty space were misattributed to the moral vignettes they judged, implicating disgust as an affective contributor to intuitive moral judgments. In another study, Wheatley and Haidt (2005) hypnotized subjects to experience disgust when they heard a key target word, such as "often". When subjects were told apparently innocent stories that contained this word—for example, a story about cousins who often visit the zoo, or a student council officer who often picks discussion topics—some made vague moral allegations against the protagonists, saying for instance "It just seems like he is up to something", or that the he is a "popularity-seeking snob."

In summary, there is research suggesting that the DDE characterizes patterns of moral intuition, and research suggesting that affective responses underlie certain moral intuitions. This raises a clear question: does contemplating harm used as the means to an end triggers an affective response that in turn generates moral judgments consistent with the DDE? We suggest that the answer is yes: although the DDE alone cannot be fashioned into a complete descriptive account of the conditions under which affective processes are engaged in moral judgments of harmful acts, there are good reasons to suppose that it captures part of the story. The DDE is among the features that distinguish "personal" from "impersonal" dilemmas used to validate the dual-process model (Greene et al., 2001; Koenigs et al., 2007; Mendez et al., 2005; Valdesolo & DeSteno, 2006). A subsequent study by Borg and colleagues (2006) also revealed increased neural activity in brain regions associated with emotion, such as the VMPC, during the judgment of cases involving the use of harm as a means to a greater good. These data suggest that the use of harm as a means may play a specific role in engaging affective processes of moral judgment; however, the stimuli used have often not been sufficiently well-controlled, leaving room for multiple interpretations.

We draw several conclusions from this collective body of data. First, the affective process that shape moral judgment often operate below the level of conscious awareness, in the sense that individuals are often unable to articulate the basis for moral judgments derived from them. Second, these unconscious mechanisms operate over representations of agents and actions, causes and intentions, etc, in ways that can be captured by fairly general principles such as the DDE. Third, although aspects of these unconscious mechanisms may involve the processing of information without emotional valence, the evidence suggests that emotion plays some causal role in generating the ultimate moral judgment. Finally, the very sorts of features that seem to guide intuitive moral judgments also seem to guide affective moral judgments. A parsimonious account of this data clearly stands out: the research programs into the affective basis of moral judgment and research programs into the intuitive basis of moral judgment have been investigating the same kind of psychological process.

3. Reasoning from deontological principles

While the research described above associates conscious, principled reasoning with characteristically consequentialist moral judgment and emotional intuition with characteristically deontological moral judgment, other evidence suggests that this pattern need not hold in all cases. Consider, first, deontological philosophy. It seems that philosophical reasoning can lead to judgments that are not consequentialist and that are even strikingly counterintuitive. (See, for example, Kant's (1785/1983) infamous claim that it would be wrong to lie to a would-be murderer in order to save someone's life.) Even when intuitions drive deontological distinctions, reasoning determines their actual role in normative theory – whether and the extent to which we should take them seriously. To adapt a dictum from evolutionary biology: intuition proposes, reason disposes.

But philosophers aren't the only ones providing evidence of non-consequentialist reasoning that appears to be conscious, principled, and deliberate. Let's return briefly to the patients with emotional deficits due to VMPC damage (Koenigs et al., 2007). The sketch we provided of their performance on personal scenarios was just that – a sketch. The full picture is both richer and messier. The personal moral scenarios on which VMPC patients produced abnormally consequentialist judgments could in fact be sub-divided into two categories: "low-conflict" or easy scenarios and "high-conflict" or difficult scenarios. Low-conflict scenarios elicited 100% agreement and fast reaction times from healthy control subjects; high-conflict scenarios elicited disagreement and slow responses. All high-conflict scenarios featured a "consequentialist" option in which harm to one person could serve to promote the welfare of a greater number of people. Low-conflict scenarios, by contrast, typically described situations in which harm to one person served another person's purely selfish ends, for example, throwing one's baby in a dumpster to avoid the financial burden of caring for it. In these cases, the VMPC patients judged the actions to be wrong, just as normal individuals do.

How do VMPC patients arrive at the 'appropriate' answer on low-conflict or easy personal moral scenarios? One proposal is that low-conflict scenarios pit a strong emotional response to the harmful action against a weak case for the alternative. According to this proposal, VMPC subjects could have generated the normal pattern of judgments on low-conflict scenarios because they retained sufficiently intact emotional processing to experience an aversion to the harm. This proposal isn't entirely plausible, however, in light of the fact that the VMPC subjects tested show abnormal processing of even highly charged emotional stimuli.

According to an alternative proposal, VMPC patients reasoned their way to conclusions against causing harm. The difference between low-conflict and high-conflict scenarios is driven by the absence of conflicting moral norms in the low-conflict scenarios and the presence of conflicting moral norms in the high-conflict scenarios. As described above, high-conflict scenarios described situations that required harm to one person to help other people. The decision of whether to endorse such harm presents participants with a moral dilemma, in the sense that they have distinct moral commitments demanding opposite behaviors. Regardless of the decision, either a norm against harming or a norm against not helping is violated. Low-conflict scenarios, on the other hand, typically described situations that required harm to one person in order to help only oneself. Thus, it is possible that an uncontested moral norm against harming

someone purely for self-benefit guides judgment. Alternatively, the VMPC patients could be applying the same utilitarian principles that they applied to the high-conflict cases, judging, for example, that the financial benefits to the young mother are outweighed by the harm done to her infant.

There were some low-conflict dilemmas featuring situations in which harm to one person was required to save other people. These scenarios, however, feature actions that violate widely understood moral norms, for instance norms against child prostitution, cannibalism, and the violation of a patient's rights by his doctor. The pattern of data thus suggests that patients with compromised emotional processing are able to use their intact capacity for reasoning to apply clear moral norms to specific situations.

Finally, ordinary people appear capable of bringing explicit moral principles to bear on particular cases in service of moral judgment. In contrast to the distinction between intended and foreseen harm, which appears to be intuitively generated and then justified *post-hoc*, the distinction between killing and letting die (or, more generally, between action and omission) may commonly be consciously deployed in the first place. Cushman and colleagues (2006) asked people to justify patterns of their own judgments that were consistent with these two distinctions. They found that many people were able to explicitly produce a version of the action/omission distinction, while very few were able to explicitly produce a version of the distinction between intended and foreseen harm. Thus, it is at least possible that ordinary people engage in moral reasoning from some version of the action/omission distinction, just as some philosophers do. Even more revealing is a reanalysis of that data showing that participants who were able to *articulate* the action/omission distinction showed significantly greater *use* of the distinction in their prior judgments. This link between having a principle and using that principle provides stronger evidence that explicit moral reasoning can play a role in the process of moral judgment.

Further evidence for the role of moral reasoning—albeit a limited role—comes from a study by Lombrozo (2008). Participants were asked several questions designed to reveal whether they adhered to a more consequentialist or more deontological moral theory. For instance, they were asked whether it is "never permissible" to murder (Lomborzo designated this a deontological response), "permissible" if it would bring about a greater good (consequentialist), or "obligatory" if it would bring about a greater good (strong consequentialist). They were also asked to make a moral judgment of a two specific cases: the "switch" and "push" variants of the trolley problem. A correlation emerged between subjects' general commitments and their judgment of particular cases, but this effect was limited in a revealing way. As predicted, subjects whose general commitments were consequentialist exhibited greater consistency in their moral judgments of the "push" and "switch" cases. This effect only emerged, however, when the "push" and "switch" cases were judged side-by-side. When specific judgments were made sequentially, there was no effect of subjects' general moral commitments. This suggests that moral reasoning can be triggered in some specific contexts, but not others.

We expect that further research will uncover further evidence for conscious, principled reasoning in moral judgments. These examples here serve to support its potential role in non-consequentialist moral judgments – adding important detail to the dual-process model.

4. The pervasive role of affect

So far we have considered the general properties of two different processes that shape moral judgment: a deliberate, effortful process that reasons about specific cases from explicit abstract principles, and a rapid, automatic process of moral judgment that generates affective responses to specific cases on the basis of mental processes inaccessible to conscious reflection. We have also begun to characterize these systems at a more computational level, specifying the content of their moral rules: A general principle favoring welfare-maximizing behaviors appears to be supported by controlled cognitive processes, while a principle prohibiting the use of harm as a means to a greater good appears to be part of the process that generates intuitive emotional responses. We conclude by turning to a crucial, unanswered question: How do these principles get into our heads?

Elsewhere Greene has suggested that some of our emotionally-driven moral judgments have an innate and evolutionarily adaptive basis (Greene & Haidt, 2002; Greene, 2003, 2007). Recent research demonstrating sophisticated social evaluation in preverbal infants has begun to lend credence to this view (Hamlin, Wynn, & Bloom, 2007). At the very least, it points the way towards research paradigms that could support a nativist hypothesis. While we look forward eagerly to progress on this front, here we won't pursue the question of the developmental or adaptive origins of the principles at work on the intuitive emotional side of the dual-process divide.

Our aim in this section is instead to consider the origins of the cognitive commitment to a utilitarian principle favoring welfare-maximizing choices (hereafter, the "welfare principle"). We explore the Humean (1739/1978) hypothesis that utilitarian judgment, despite its association with controlled cognition (Greene et al., 2004, in press) and its prominence in the presence of emotional deficits (Mendez et al., 2005; Koenigs et al., 2007, Ciaramelli et al., 2007), itself has an affective basis. Put simply, we suggest that affect supplies the primary motivation to regard harm as a bad thing, while 'cognition' uses this core statement of value to construct the utilitarian maximum that we ought to act so as to minimize harm. The result is the welfare principle. Below, we sketch out two versions of this hypothesis in greater detail.

They begin with a proposal by Greene (2007) distinguishing two kinds of emotional responses: those that that function like alarm bells and those that function like currencies. This distinction is meant to capture the difference between deontological intuitions and utilitarian reasoning, but it is also offered as a broader distinction that operates beyond the moral domain. The core idea is that alarm-bell emotions are designed to circumvent reasoning, providing absolute demands and constraints on behavior, while currency emotions are designed to participate in the process of practical reasoning, providing negotiable motivations for and against different behaviors. For example, the amygdala, which has been implicated in responses to personal moral dilemmas, reliably responds to threatening visual stimuli such as snakes and faces of out-group members (LeDoux, 1996: 166; Phelps et al., 2000). Thus, the amygdala is a good candidate for a region that is critical for supporting at least some alarm-bell emotions. In contrast, Knutson and colleagues (1995) have identify a set of meso-limbic brain regions that appear to represent expected monetary value in a more graded fashion, with distinct

regions tracking a stimulus' reward magnitude, reward probability, and expected value. These regions, in a rather transparent way, support currency-like representations.

Above, we argued that strongly affective deontological intuitions are triggered when types of harms occur, perhaps those that involve physical contact and use harm as a means to accomplishing a goal. The emotional response is like an alarm bell because it is designed to make a clear demand that is extremely difficult to ignore. This response can be overridden with substantial effort (cognitive control), but it is not designed to be negotiable. The prohibition of certain physical harms seems to be just example of a broader class of alarm bell emotions. For instance, research on 'protected values' (Baron & Spranca 1997) and 'sacred values' (Tetlock, Kristel, Elson, Green & Lerner 2000) in moral judgment suggest that people sometimes treat harm to the environment as non-negotiable as well. Outside the moral domain, the strong disgust response to eating contaminated food also seems to have an alarm-like character: most people are not willing to negotiate eating feces.

Whereas alarm bell emotions treat certain types of physical harms as non-negotiable, it is precisely the defining feature of the 'welfare principle' that harms are negotiable. One person's life counts for one-fifth as much as five people's lives, and the relative costs and benefits of lives, among other outcomes, are weighed against each other to determine the best course of action. Here, again, the case of moral decision-making is presumed to be one case of a much broader phenomenon. The desire for ice cream on a hot summer day is an example of a currency emotion: it supplies a reason to pursue the Good Humor truck, but this reason can be traded off against others, such as maintaining a slim poolside profile. Currency-like emotions function by adding a limited measure of motivational weight to a behavioral alternative, where this weighting is designed to be integrated with other weightings in order to produce a response. Such emotional weightings, rather than issuing resolute commands, say, "Add a few points to option A" or "Subtract a few points from Option B."

Above, we suggested that affect supplies the primary motivation to regard harm as bad. Once this primary motivation is supplied, reasoning proceeds in a currency-like manner. But we still must explain the origin of the primary motivation to regard harm as bad . Is it located in the alarm bell response to personal harms, or in an independent currency response? Each hypothesis has pros and cons.

According to the 'alarm bell' hypothesis, the primary motivation not to harm is ultimately derived from the alarm bell emotional system that objects to things like pushing a man in front of a train. When people attempt to construct general principles that account for their particular 'alarm bell' moral intuitions, one of the first things they notice is that their intuitions respond negatively to harm. This gives rise to a simple moral principle: "harm is bad". Combined with a general cognitive strategy of minimizing bad things (i.e., practical reasoning), the result is the welfare principle. Represented as such, the welfare principle can become a basis for controlled, conscious processes of reasoning. Critically, reasoning on the basis of an explicit welfare principle can occur in a currency-like manner. The alarm-bell hypothesis locates the *origins* of the welfare principle in a process of theory-construction over alarm-bell responses, but maintains the view that the *operation* of the welfare principle occurs in a currency-like manner that engages controlled reasoning processes.

13

According to this hypothesis, the welfare principle takes hold not because it offers a fully adequate descriptive account of our intuitive moral judgments (which it does not), but because it is simple, salient, and accounts for a large proportion of our intuitive judgments (which it does). Ultimately, the same mechanism can also gives rise to more complex moral principles. For instance, Cushman and colleagues (in preparation) have explored this hypothesis in the particular case of the doctrine of double effect and doctrine of doing and allowing.

The principle virtue of the alarm bell hypothesis is its parsimony: by explaining the origins of the welfare principle in terms of an independently-hypothesized alarm bell aversion to harm, it can provide a motivational basis for controlled cognitive moral reasoning without invoking any additional affective mechanisms. It can also explain why utilitarian moral judgment is preserved in individuals who experience damage to frontal affective mechanisms: the welfare principle has already been constructed on the basis of past affective responses. But one shortcoming of the alarm bell hypothesis is that it leaves unexplained how a theory of one's own moral intuitions gives rise to practical reasoning. When an individual regards her own pattern of moral intuitions and notes, "I *seem to think* harm is bad", will this lead automatically to the conclusion "Harm *is a reason* not to perform a behavior"? At present, we lack a sufficient understanding of the interface between cognition and affect to answer this difficult question. It seems plausible that a descriptive theory of one's motivations could become a motivational basis for practical reasoning, but it also seems plausible that it might not.

Philosophers will note that the alarm bell hypothesis paints an unflattering portrait of philosophical utilitarianism because it characterizes the welfare principles as a sort of crude first pass, while characterizing deontological principles as more subtle and sophisticated. People's intuitive judgments are often consistent with the welfare principle, but it is clear that in many cases they are not—for instance, in the footbridge version of the trolley problem. If the welfare principle is designed capture our intuitive moral judgments as a whole, then it would receive about a B+, getting things descriptively right much of the time, but getting things descriptively wrong much of the time, too.

The currency hypothesis is more friendly toward utilitarianism/consequentialism as a normative approach. According to the currency hypothesis, we are furnished with certain currency-like affective responses *independently* of our alarm bell responses. For instance: Harm is bad, regardless of who experiences it. Benefits are good, regardless of who experiences them. More harm is worse than less harm. More benefits are better than fewer benefits. Small harms can be outweighed by large benefits. Small benefits can be outweighed by large harms. And so on.

Notice that these premises assign a valence to outcomes (harm is bad), and even ordinal relationships among outcomes (more harm is worse), but they do not assign cardinal values to particular outcomes. Thus, they specify the general structure of utilitarian thinking, but do not specify how, exactly, various harms and benefits trade off against one another. According to the currency process, utilitarian thinking is the product of a rational attempt to construct a set of practical principles that is consistent with the constraints imposed by the aforementioned premises. It is an idealization based on the principles that govern the flow of emotional currency. One might say that it is a union between basic sympathy and basic math.

14

Note that the operational mechanisms upon which the currency hypothesis depends—the math, so to speak—overlap substantially with the mechanisms upon which the alarm bell hypothesis depends. Both accounts posit that explicit utilitarian principles arise from a combination of abstract reasoning and affect. The key difference is the source of the affect: generalization from alarm-bell responses prohibiting harm, versus an independent set of currency-like valuations of harms and benefits.

The principle virtue of the currency hypothesis is that utilitarian cost/benefit reasoning looks very much like cost/benefit reasoning over other currency-like responses. The welfare principle functions very much as if there were a negotiable negative value placed on harm—and, for that matter, a negotiable positive value placed on benefits. Also, in contrast to the alarm bell hypothesis, it is apparent how the currency-like weighting of costs and benefits directly and necessarily enters into practical reasoning. This tight fit comes with a slight expense in parsimony, however. The currency hypothesis demands two separate types of affective response to harm of independent origin: an alarm bell response to a select set of harms, and a currency response to a larger set of harms. It also implies that currency-like responses are preserved in individuals with frontal-lobe damage, since they continue to reason from the welfare principle.

As noted earlier, the currency hypothesis is more friendly toward philosophical utilitarianism. According to this view, utilitarians are not simply doing a poor job of generalizing over the body of their alarm bell moral intuitions. Instead, their judgments are based indirectly on a distinct set of currency-like emotional responses. Is it better to rely on currency-like emotions to the exclusion of alarm-like emotions? Perhaps. The alarm-like emotions that drive people's non-utilitarian judgments in response to trolley dilemmas appear to be sensitive to factors that are hard to regard as morally relevant, such as whether the action in question involves body-contact between agent and victim (Cushman et al., 2006). Taking the charge of bias one step further, Greene (in preparation) hypothesizes that the DDE is a by-product of the computational limitations of the processes that govern our intuitive emotional responses. Borrowing some computational machinery from Mikhail (2000), he argues that the controlled cognitive system based in the DLPFC has the computational resources necessary to represent the side-effects of actions, while the appraisal system that governs our emotional responses to actions like the one in the footbridge dilemma lacks such resources. As a result, our emotional responses have a blind spot for harmful side-effects, leading us to draw a moral distinction between intended and foreseen harms, i.e. the DDE.

Although the alarm bell and currency hypotheses vary in detail, they both reject philosophical moral rationalism in that they (a) require a specification of primitive goods before practical reasoning (including utilitarian reasoning) can proceed and (b) locate these primitives in our affective responses. Moreover, these hypotheses are not mutually exclusive: the welfare principle may be supported both by theory-building based on alarm bell responses as well as a distinct set of currency responses. As we argued above, there is ample evidence that a distinction between cognitive and affective processes of moral judgment is warranted. We strongly suspect, however, that when the origins of the cognitive process are understood, we will find a pervasive influence of affect.

Conclusion

As we hope this essay attests, it no longer makes sense to engage in debate over whether moral judgment is accomplished exclusively by "cognition" as opposed to "affect", or exclusively by conscious reasoning as opposed to intuition. Rather, moral judgment is the product of complex interactions between multiple psychological systems. We have focused on one class of moral judgments: those involving tradeoffs between avoiding larger harms and causing smaller ones. Cases such as these engage at least two distinct systems: an intuitive/affective response prohibiting certain kinds of basic harm, and a conscious/cognitive response favoring the welfare-maximizing response. The underlying psychology of moral judgment in these cases helps to explain why they strike us as particularly difficult moral dilemmas: we are forced to reconcile the conflicting output of competing brain systems.

We have also identified several aspects of this account that are in need of further investigation. First, there is much to be learned about the evaluative processes that operate within each of the systems we have identified. In the case of the intuitive/affective system, we have suggested that one component of the evaluative process mirrors the doctrine of double effect. But the evidence is not conclusive on this point. Moreover, this single principle alone cannot account for the full pattern of data associated with intuitive/affective moral judgments. In the case of the conscious/cognitive system, the data strongly suggests that ordinary people typically reason from a principle favoring welfare-maximizing choices. But there is also good reason to believe that people, at least in some circumstances, explicitly reason from deontological moral principles. Further research in this area will be critical.

Second, the origins of each system of moral judgment remain unknown. In this essay we have explored two hypotheses concerning the origins of explicit moral principles within individuals, both of which ultimately implicate affective systems. Others have proposed diverse origins for moral principles such as metaphor (Lakoff & Johnson, 1999: 290-334) and construction from social experience (Kohlberg, 1969). A further question is the origin of moral intuitions, which several authors have suggested may have an innate basis (Hauser, 2006: 1; Mikahil, 2007; Greene & Haidt, 2002). Several of these proposals operate at distinct levels of analysis (ontogenetic versus evolutionary) and are specific to one of the two systems we have discussed (cognitive versus affective), and thus may be broadly compatible.

Third, at present we know little about how the intuitive/affective and conscious/cognitive systems interact on-line in the production of moral judgments. This is a topic we have left largely untouched in the present essay, and somewhat out of necessity. Until the contours of individual systems of moral judgment are better understood, it will be difficult to make much progress towards understanding the interactions between systems. One aspect of this problem that we suspect will be of interest are the standards for the normative plausibility of putative moral principles. Certain factors that reliably shape moral judgments, such as the physical proximity of an agent to a victim, are sometimes rejected as *prima face* invalid criteria for moral judgment. Others, such as the doctrine of double effect, are more commonly accepted (Cushman et al., 2006). The bases on which particular explicit moral rules are accepted or rejected are poorly understood, and such meta-ethical intuitions are a key area for investigation at the interface between the intuitive/affective and conscious/cognitive systems.

Finally, it remains to be seen how generally the dual-process model developed for harm-based moral dilemmas can be extended to other domains of morality. There are at least two ways that this issue can be framed. On the one hand, we have argued that in the specific case of tradeoffs in harms, conflict between distinct psychological systems gives rise to the phenomenon of a dilemma (Cushman & Young, in press; Greene, 2008; Sinnott-Armstrong, in preparation). One important question is whether this multi-system account can be employed to understand the phenomenon of a moral dilemma beyond the domain of physically harmful actions. That is, are there distinct systems that give rise to potentially conflicting moral judgments in domains such as the division of economic resources, the establishment of conventional standards of conduct, sexual taboo, and so forth?

On the other hand, we have argued for the operation of two psychological processes: an intuitive/affective response to intentional harms and a conscious/cognitive response favoring welfare maximization. This raises a second question: To what extent do other types of moral judgment depend on the operation of these specific processes? For instance, when evaluating allocations of resources, financial losses could be coded as "harms" and then processed via the operation of one of the systems we have explored in this essay. Whether these particular systems have such broad applicability is presently unknown.

It is, of course, our hope that the small corner of moral judgment we have explored in this essay—a corner strictly limited to the occurrence of physical harms and preternaturally concerned with railway operations—will teach lessons with broad applicability. The extent of this applicability remains to be determined, but we feel confident in asserting at least this much: There is no single psychological process of moral judgment. Rather, moral judgment depends on the interaction between distinct psychological systems. Notably, these systems tend to be reflected in the principles characteristic of competing moral philosophies (Greene 2007).

References

Baron, J. & Spranca, M. (1997). Protected values. *Behavior and human decision processes*, 70, 1-16.

Bazerman, M., Loewenstein, G., & Blount White, S. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. Administrative Science Quarterly, 37(220-240).

Bjorklund, F., Haidt, J., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. Lund Psychological Reports, 2, 1-23.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. Cognition, 57, 1-29.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. (2001). Conflict monitoring and cognitive control. Psychological Review, 108(3), 624-652.

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

Cushman, F. A., & Young, A. W. (in press). The psychology of dilemmas and the philosophy of morality. Ethical Theory and Moral Practice.

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.

Cushman, F. A., Young, L., & Hauser, M. D. (in preparation). Patterns of moral judgment derive from non-moral psychological representations.

Damasio, A. (1994). Descartes' Error. Boston, MA: Norton.

Darley, J. M., & Shultz, T. R. (1990). Moral Rules - Their Content And Acquisition. Annual Review of Psychology, 41, 525-556.

Fischer, J. M., & Ravizza, M. (1992). Ethics: Problems and Principles. New York: Holt, Rinehart & Winston.

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

Gilligan, C. (1982/1993). In a Different Voice: Psychological Theory and Women's Development. Cambridge: Harvard University Press.

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), Moral Psychology (Vol. 3). Cambridge: MIT Press.

Greene, J. D., Cohen, J. D., Nystrom, L. E., Lindsell, D., Clarke, A. C., & Lowenberg, K. (in prep). What pushes your moral buttons?: Modular mypopia and the trolley problem.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Science, 6, 517-523.

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.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814-834.

Haidt, J. (2007). The new synthesis in moral psychology. Science, 316(5827), 998-1002.

Hamlin, K., Wynn, K., & Bloom, P. (2008). Social evaluation by preverbal infants. Nature, 450, 557-559.

Hauser, M. D. (2006). Moral Minds: How nature designed a universal sense right and wrong. New York: Harper Collins.

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.

Huebner, B., Dwyer, S., & Hauser, M. (2008). The role of emotion in moral psychology. Trends in Cognitive Science, 13(1)

Hume, D. (1739/1978). A treatise of human nature. In L. Selby-Bigge & P. H. Nidditch (Eds.).

Kamm, F. M. (1998). Morality, Mortality: Death and whom to save from it. New York: Oxford University Press.

Kamm, F. M. (2006). Intricate ethics: Rights, responsibilities, and permissible harm. New York: Oxford University Press.

Kant, I. (1785/1983). On a supposed right to lie because of philanthropic concerns. Indianapolis: Hackett.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *J Neurosci, 25*(19), 4806-4812.

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: Academic Press.

LeDoux, J. (1996). *The Emotional Brain*. New York: Simon & Schuster.

Lombrozo, T. (in press). The role of moral commitments in moral judgment. *Cognitive Science.*

Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh. New York: Basic Books.

Macmillan, M. (2000). An odd kind of fame. Cambridge: MIT Press.

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 Science, 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'. Unpublished PhD, Cornell University, Ithaca.

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

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *J Cogn Neurosci, 12*(5), 729-738.

Piaget, J. (1965/1932). The Moral Judgment of the Child. New York: Free Press.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. Journal of Experimental Social Psychology, 39(653-660).

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S. T., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. Journal of Cognitive Neuroscience, 18(5), 803-837.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (in press). Disgust as embodied moral judgment. Personality and Social Psychology Bulletin.

Schweder, D., & Haidt, J. (1993). The future of moral psychology: truth, intuition, and the pluralist way. Psychological Science, 4, 360-365.

Sinnott-Armstrong, W. (in prep). Abstract + Concrete = Paradox.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 12, 643–662.

Tetlock, P., Kristel, O., Elison, S., Green, M. & Lerner, J. (2000). The psychology of the unthinkable. Journal of Personality and Social Psychology, 78(5) 853-870.

Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 279.

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

Turiel, E. (2005). Handbook of Moral Development. In M. Killen & J. Smetana (Eds.), Handbook of Moral Development: Lawrence Erlbaum Associates Publishers.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. Psychological Science, 17(6), 476-477.

Waldman, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. Psychological Science, 18(3), 247-253.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. Psychological Science, 16(10), 780-784.