The Cognitive Neuroscience of Moral Judgment

Joshua D. Greene

To appear in

*The Cognitive Neurosciences IV*

Department of Psychology
Harvard University
33 Kirkland St.
Cambridge, MA  02138
617-495-3898
jgreene@wjh.harvard.edu
This article reviews recent advances in the cognitive neuroscience of moral judgment. The field began with studies of individuals who exhibit abnormal moral behavior, including neurological patients and psychopaths. Such studies continue to provide valuable insights, particularly concerning the role of emotion in moral decision-making. Recent functional neuroimaging studies of normal individuals have identified neural correlates of specific emotional processes relevant to moral judgment. A range of studies using diverse methods support a dual-process theory of moral judgment according to which utilitarian moral judgments (favoring the “greater good” over individual rights) are enabled by controlled cognitive processes, while deontological judgments (favoring individual rights) are driven by intuitive emotional responses. Several recent neuroimaging studies focus on the neural bases of mental state attribution in the context of moral judgment. Finally, research in the field of neuroeconomics has focused on neural processing related to cooperation, trust, and fairness.
The aim of cognitive neuroscience is to understand the mind in physical terms. This endeavor assumes that the mind can be understood in physical terms, and, insofar as it is successful, validates that assumption. Against this philosophical backdrop, the cognitive neuroscience of moral judgment takes on special significance. Moral judgment is, for many, the quintessential operation of the mind beyond the body, the Earthly signature of the soul (Greene, in press). (In many religious traditions it is, after all, the quality of a soul’s moral judgment that determines where it ends up.) Thus, the prospect of understanding moral judgment in physical terms is especially alluring, or unsettling, depending on your point of view.

In this brief review I provide a progress report on our attempts to understand how the human brain makes moral judgments. In recent years, we have continued to learn valuable lessons from individuals whose abnormal brains dispose them to abnormal social behavior. We have developed new moral-psychological testing materials and used them to dissociate and characterize the affective and cognitive processes that shape moral decisions. Finally, the field of neuroeconomics has brought a welcome dose of ecological validity to the study of moral decision-making. I discuss each of these developments below. (Important and relevant developments in other fields, such as animal behavior and developmental psychology (de Waal, 2006; Hamlin et al., 2007; Warneken et al., 2007), are beyond the scope of this article.)
Bad brains

In the 1990s, Damasio and colleagues published a series of path-breaking studies of decision-making in patients with damage to ventromedial prefrontal cortex (VMPFC), one of the regions damaged in the famous case of Phineas Gage (Damasio, 1994; Macmillan, 2000). VMPFC patients were mysterious because their real-life decision-making was clearly impaired by their lesions, but their deficits typically evaded detection using standard neurological measures of executive function (Saver and Damasio, 1991). Notably, such patients showed no sign of impairment on Kohlberg’s (Colby and Kohlberg, 1987) widely used test of moral reasoning (Anderson et al., 1999). Using a game designed to simulate real-world risky decision-making (The Iowa Gambling Task), Bechara and colleagues (1996) documented these behavioral deficits and demonstrated, using autonomic measures, that these deficits are emotional. It seems that such patients make poor decisions because they are unable to generate the feelings that guide adaptive decision-making in healthy individuals.

A later study targeting moral judgment (Anderson et al., 1999) compared patients with adult-onset VMPFC damage to two patients who had acquired VMPFC damage as young children. While the late-onset patients make poor real-life decisions (e.g. neglecting relatives and friends, involvement in shady
business ventures), indicating a deterioration of “moral character” (Damasio, 1994), their behavior tended to harm themselves as much as others. The early-onset patients, however, developed into “sociopathic” adults who, in addition to being irresponsible and prone to risk-taking, are duplicitous, aggressive, and strikingly lacking in empathy. What’s more, these two patients, unlike the late-onset patients, exhibited a child-like “preconventional” pattern of moral judgment, reasoning about moral issues from an egocentric perspective focused on reward and punishment. This result suggests a critical role for emotion in moral development. The late-onset patients are prone toward bad decision-making, but, thanks to a lifetime of emotional experience, they are not truly sociopathic. The early-onset patients, in contrast, lacked the emotional responses necessary to learn the basics of human moral behavior. (See also Grattan & Eslinger (1992))

Studies of psychopaths and other individuals with anti-social personality disorder (APD) underscore the importance of emotion in moral decision-making. APD is a catch-all diagnosis for individuals whose behavior is unusually anti-social. Psychopathy, in contrast, is a more specific, somewhat heritable (Blonigen et al., 2005; Viding et al., 2005) disorder whereby individuals exhibit a pathological degree of callousness, lack of empathy or emotional depth, and lack of genuine remorse for their anti-social actions (Hare, 1991). Psychopaths tend to engage in instrumental aggression, while other individuals with APD are characterized by reactive aggression (Berkowitz, 1993; Blair, 2001).
Psychopathy is characterized by profound, but selective emotional deficits. Psychopaths exhibit normal electrodermal responses to threat cues (e.g. a picture of shark’s open mouth), but reduced responses to distress cues (e.g. a picture of a crying child) (Blair et al., 1997). In a particularly revealing study, Blair (1995) demonstrated that psychopaths fail to distinguish between rules that authorities cannot legitimately change (“moral” rules, e.g. a classroom rule against hitting) from rules that authorities can legitimately change (“conventional” rules, e.g. a rule prohibiting talking out of turn). According to Blair, psychopaths see all rules as mere rules because they lack the emotional responses that lead ordinary people to imbue moral rules with genuine, authority-independent moral legitimacy.

Findings concerning the specific neural bases of psychopathy and APD are varied, implicating a wide range of brain regions including the orbital frontal cortex (OFC)/VMPFC, insula, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), amygdala, parahippocampal gyrus, and superior temporal gyrus (Kiehl, 2006; Raine and Yang, 2006; Muller et al., 2008). Blair (2004; 2007) has proposed that psychopathy arises primarily from amygdala dysfunction, which is crucial for stimulus-reinforcement learning (Davis and Whalen, 2001), and thus for normal moral socialization (Oxford et al., 2003). The amygdala exhibits reduced activity in psychopaths both in non-moral contexts (e.g., in response to emotional words (Kiehl et al., 2001)) as well in socio-moral contexts (e.g. during cooperative behavior in a prisoner’s dilemma game (Rilling
et al., 2007)). Consistent with this view, Yang and colleagues (2006) found reduced amygdala volume in psychopaths. The VMPFC, which is known to work in concert with the amygdala (Diergaarde et al., 2005; Schoenbaum and Roesch, 2005), also exhibits many of these effects and appears to play a role in (mis)representing the value of behavioral outcomes in psychopathy (Blair, 2007).

A broader suite of brain regions have been implicated in APD (Raine and Yang, 2006), suggesting, among other things, more general deficits in prefrontal function (Raine et al., 1994). These may be due to structural abnormalities involving reduced prefrontal gray matter (Raine et al., 2000; Yang et al., 2005). There is some evidence implicating abnormal function in dorsolateral prefrontal cortex (DLPFC) in patients with APD (Schneider et al., 2000; Vollm et al., 2004), but not in patients with psychopathy. Given the DLPFC’s role in cognitive control (Miller and Cohen, 2001), this is consistent with the notion that psychopaths’ aggression results from lack of empathy for others, rather than poor impulse control.

**Mapping moral emotion**

Consistent with research on APD, and in keeping with a broader trend in moral psychology (Haidt, 2001), most research using functional imaging to study morality has focused on mapping the “where” and “when” of moral emotion in the brain. Some early studies compared moral and non-moral stimuli (Moll et
al., 2001; Moll et al., 2002a; Moll et al., 2002b) and identified a suite of brain regions that are sensitive to moral stimuli including the OFC, mPFC, frontal pole, PCC/precuneus, superior temporal sulcus (STS), and temporal pole. This approach, while informative, depends critically on the choice of non-moral control stimuli and the assumption that the observed results are in some way specific to morality (Greene and Haidt, 2002). More recent functional imaging studies have focused on identifying and functionally characterizing different kinds of moral-emotional processes.

*Empathy, caring, and harm:* Greene and colleagues (2001) identified a set of brain regions associated with judging actions involving “personal,” as compared to “impersonal,” harm: mPFC (BA 9/10), the PCC/Precuneus (BA 31), and the posterior superior temporal sulcus (pSTS)/temperoparietal junction (TPJ)/angular gyrus (BA 39). (See Figure 1 and “Dual-process morality” below for more details.) A study replicating these results using a larger sample (Greene et al., 2004) identified the same effect in the amygdala, among other regions. The aforementioned regions are implicated in emotional processing (Maddock, 1999; Phan et al., 2002; Adolphs, 2003) as well as in “theory of mind” (ToM) (Frith, 2001; Adolphs, 2003; Amodio and Frith, 2006; Young et al., 2007). These regions, with the exception of the amygdala, are also part of the “default network,” (Gusnard and Raichle, 2001; Fox et al., 2005), a set of brain regions that exhibits relatively high levels of tonic activity and that reliably decreases in activity.
during outwardly directed tasks. Parts of this network are also implicated in self-referential processing (Gusnard et al., 2001; Kelley et al., 2002), episodic memory, “prospection” (Buckner and Carroll, 2007; Schacter et al., 2007), and mind-wandering (Mason et al., 2007). A persistent theme among these processes is the representation of events beyond the observable here and now, such as past, future, and imagined events and mental states. Thus, the activity observed in this network during the contemplation of dilemmas involving “personal” harm is probably related to the fact that these stimuli involve such non-sensory representations, although this alone does not explain why “personal” dilemmas engage this network more than “impersonal” ones. Consistent with this idea, the functional imaging studies of moral judgment that have most robustly engaged this network involve more complex, text-based, narrative stimuli (Greene et al., 2001; Greene et al., 2004; Schaich Borg et al., 2006; Robertson et al., 2007; Young et al., 2007; Greene et al., 2008b; Greene et al., 2008d; Schaich Borg et al., 2008; Young and Saxe, 2008; Kedia et al., in press).

Several studies have focused on neural responses to different types of harm. Luo and colleagues (2006) found that the right amygdala and left VMPFC are sensitive to the intensity of harm displayed in pictures, and Heekeren et al. (2005) found that the amygdala exhibits increased activity in response to narratives involving bodily harm. An earlier study (Heekeren et al., 2003) found no effects in the amygdala using stimuli devoid of violence. Finally, individuals with high psychopathy scores exhibited decreased amygdala activity during the
contemplation of moral dilemmas involving “personal” harm (Glenn et al., 2008). Thus, the evidence from functional imaging suggests that the amygdala plays an important role in triggering emotional responses to physically harmful actions.

*Specific emotions:* Several studies of moral judgment have focused on specific moral emotions, including moral disgust (Rozin et al., 1994; Rozin et al., 1999; Wheatley and Haidt, 2005). Moll and colleagues (Moll et al., 2005) identified a number of brain regions sensitive to core/pathogen disgust in moral contexts. A more recent study (Schaich Borg et al., 2008) compared disgust in response to incestuous acts to pathogen disgust and moral transgressions of a non-sexual nature. Stimuli describing non-sexual moral transgressions (e.g. lying, cheating, stealing), as compared to pathogen-disgust stimuli, elicited increased activity in the familiar mPFC/PCC/TPJ network, but also in the frontal pole/DLPFC and the ACC. Incest descriptions, as compared to that of non-sexual moral transgressions, elicited increased activity in the mPFC/PCC/TPJ network and other regions, including the inferior frontal gyrus, the left insula, the ventral and dorsal ACC, the left amygdala, and the caudate nucleus. Perhaps surprisingly (Phillips et al., 1997; Calder et al., 2001), the insula was preferentially activated only in the incest condition.

Other studies have focused on social emotions such as guilt, embarrassment, shame, pride, empathy, and anger. Robertson and colleagues (2007) found that stimuli associated with care-based morality, as compared to
justice-based morality, elicited greater activity in the mPFC and OFC, while the reverse effect was observed in the intraparietal sulcus. In a notably clever study, Beer and colleagues (2003) observed that patients with OFC damage exhibited inappropriate lack of embarrassment when given an opportunity to disclose personal information, inappropriate embarrassment when over-praised for an unremarkable performance on a simple task, and inappropriate pride and lack of embarrassment when describing the nickname they had invented for the experimenter. Berthoz and colleagues (2006) found that the amygdala is especially responsive to the evaluation of intentional transgressions committed (hypothetically) by oneself (guilt), while Kedia and colleagues (in press) observed increased activity in the mPFC, precuneus, and TPJ for evaluations of transgressions involving others (guilt, anger, compassion). These researchers also observed increase activity in the amygdala, ACC, and basal ganglia for transgressions in which both the self and another are involved (guilt, anger). (See also Shin et al. (2000), Berthoz et al. (2002), and Takahashi et al. (2004)).

Moral emotion in context: Other studies have examined the contextual modulation of moral emotion. King and colleagues (2006) used a custom-designed video game to examine violent vs. compassionate behavior in situations in which the behavior is either appropriate (harming an aggressive enemy, helping a distressed innocent person) or inappropriate (helping an aggressive enemy, harming a distressed innocent person). They found that appropriate behavior
(whether violent or compassionate) was associated with increased activity in the amygdala and VMPFC. Harenski and Hamann (2006) found that subjects who consciously down-regulated their moral emotions in several lateral regions of PFC. Finally, Finger and colleagues (2006) found that stimuli describing moral/social transgressions committed in the presence of an audience elicited increased activity in the amygdala, underscoring the importance of the amygdala in the social regulation of transgressive behavior (Blair, 2007).

**Dual-process morality**

The research described above emphasizes the role of emotion in moral judgment (Haidt, 2001), while traditional theories of moral development emphasize the role of controlled cognition (Kohlberg, 1969; Turiel, 2006). I and my collaborators have developed a dual-process theory (Kohlberg, 1969; Posner and Snyder, 1975; Chaiken and Trope, 1999; Lieberman et al., 2002; Kahneman, 2003) of moral judgment that synthesizes these perspectives. (See Figure 2) According to this theory, both intuitive emotional responses and more controlled cognitive responses play crucial and, in some cases, mutually competitive roles. More specifically, this theory associates controlled cognitive processing with utilitarian (or consequentialist) moral judgment aimed at promoting the “greater good” (Mill, 1861/1998). In contrast, this theory associates intuitive emotional
processing with deontological judgment aimed at respecting rights, duties, and obligations (Kant, 1785/1959) that may trump the greater good.

We developed this theory in response to a longstanding philosophical puzzle known as the Trolley Problem (Foot, 1978; Thomson, 1985; Fischer and Ravizza, 1992): First, consider the following moral dilemma (which we’ll here call the switch case (Thomson, 1985)): A runaway trolley is about to run over and kill five people, but you can save them by hitting a switch that will divert the trolley onto a side track, where it will run over and kill only one person. Here, most people say that it is morally acceptable to divert the trolley (Petrinovich et al., 1993), a judgment that accords well with the utilitarian perspective emphasizing the greater good. In the contrasting footbridge dilemma, a runaway trolley once again threatens five people. Here, the only way to save the five is to push a large person off a footbridge and into the trolley’s path, stopping the trolley but killing the person pushed. (You’re too small to stop the trolley yourself.) Here, most people say that it’s wrong to trade one life for five, consistent with the deontological perspective, according to which individual rights often trump utilitarian considerations.

We hypothesized that people tend to disapprove of the action in the footbridge dilemma because the harmful action in that case, unlike the action in the switch case, elicits a prepotent negative emotional response that inclines people toward disapproval (Figure 2e). We hypothesized further that people tend to approve of the action in the switch case because, in the absence of a
countervailing prepotent emotional response, they default to a utilitarian mode of reasoning that favors trading one life for five (Figure 2a). We proposed that the negative emotional response elicited by the *footbridge* case is related to the more “personal” nature of the harm in that case. We proposed, in other words, there is an emotional appraisal process (Scherer et al., 2001) that distinguishes personal dilemmas like the *footbridge* case from impersonal dilemmas like the *switch* case (Figure 2d).

To test these hypotheses we devised a set of “personal” dilemmas modeled loosely on (and including) the *footbridge* case and a contrasting set of “impersonal” dilemmas modeled loosely on (and including) the *switch* case. The effects of these stimuli were compared using fMRI. As predicted, the personal dilemmas preferentially engaged brain regions associated with emotion, including the mPFC, PCC, and the amygdala (Greene et al., 2001, 2004). (As noted above, this contrast also revealed preferential engagement of the pSTS/TPJ). Also consistent with our dual-process theory, the impersonal moral dilemmas, relative to “personal” ones, elicited increased activity in regions of DLFPC associated with working memory (Cohen et al., 1997; Smith and Jonides, 1997) and cognitive control (Miller and Cohen, 2001).

According to the dual-process theory, the *footbridge dilemma* elicits a conflict between utilitarian reasoning and emotional intuition, where the latter tends to dominate. In other cases these opposing forces appear to be more balanced. Consider the *crying baby* dilemma: It's 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?

Here, people are relatively slow to respond and exhibit no consensus in their judgments (Greene et al., 2004). According to the dual-process theory, these behavioral effects are the result of the aforementioned conflict between emotional intuition and controlled cognition. This theory makes two key predictions. First, if dilemmas like crying baby elicit response conflict, then we would expect these dilemmas (as compared to personal dilemmas that elicit shorter RTs and less disagreement) to be associated with increased activity in the ACC, a region known for its sensitivity to response conflict (Botvinick et al., 2001). (See Figure 2c) Second, if making utilitarian judgments in such cases requires overriding a prepotent, countervailing emotional response, then we would expect such judgments to be associated with increased activity in regions of DLPFC associated with cognitive control (Greene et al., 2001, 2004). (See Figure 2b.) Both of these predictions were confirmed (Greene et al., 2004).

Three more recent studies support the dual-process theory by indicating a causal relationship between emotional responses and deontological/non-utilitarian moral judgments. Mendez et al. (2005) found that patients with
frontotemporal dementia, who are known for their “emotional blunting,” were disproportionately likely to approve of the action in the footbridge dilemma. Koenigs et al. (2007) and Ciaramelli et al. (2007) observed similar results in patients with emotional deficits due to VMPFC lesions. The results of the former study, which distinguished high-conflict personal dilemmas such as the crying baby dilemma from low-conflict personal dilemmas, were particularly dramatic. (See Figure 3.) In each of the high-conflict dilemmas, the VMPFC patients gave more utilitarian judgments than the control subjects. Finally, Valdesolo & DeSteno (2006) found that normal participants were more likely to approve of the action in the footbridge dilemma following a positive emotion induction aimed at counteracting negative emotional responses.

Four others studies support the link between utilitarian judgment and controlled cognition. My colleagues and I conducted a cognitive load study in which subjects responded to high-conflict personal dilemmas while performing a secondary task (detecting presentations of the number “5” in a stream of numbers) designed to interfere with controlled cognitive processes. The cognitive load manipulation slowed down utilitarian judgments, but had no effect on RT for deontological/non-utilitarian judgments, consistent with the hypothesis that utilitarian judgments, unlike deontological judgments, are preferentially supported by controlled cognitive processes (Greene et al., 2008a). Three other studies have examined the relationship between moral judgment and individual differences in cognitive style. Bartels (2008) found that
individuals who are high in “need for cognition” (Cacioppo et al., 1984) and low on “faith in intuition” (Epstein et al., 1996) were more utilitarian. Along similar lines, Hardman (2008) examined moral judgment using the Cognitive Reflection Test (CRT) (Frederick, 2005), which asks people questions like this: “A bat and a ball cost $1.10. The bat costs one dollar more than the ball. How much does the ball cost?” The intuitive answer is 10¢, but a moment’s reflection reveals that the correct answer is 5¢. The people who correctly answered these questions were about twice as likely to give utilitarian responses to the footbridge and crying baby dilemmas. Finally, Moore and colleagues (2008) found that individuals with higher working memory capacity were more likely to give utilitarian judgments in response to dilemmas in which harm to the victim is inevitable. Note however that Kilgore et al. (2007) found that sleep-deprivation made subjects more utilitarian. This effect, however, was not observed in individuals high in emotional intelligence, suggesting the operation of complex emotion-cognition interactions that are not readily explained by current theory.

Three more recent fMRI studies support and broaden the dual-process theory. My colleagues and I compared dilemmas like the switch case to similar dilemmas in which saving the five requires breaking a promise. (E.g., the agent had promised the potential victim that he will not be run over.) In these cases it is the harm’s social structure, rather than its physical structure, that generates the tension between utilitarian and deontological judgment. We found, first, that introducing the promise factor reproduces the familiar pattern of
mPFC/PCC/TPJ activity and, second, that utilitarian judgment in the promise dilemmas is associated with increased activity in the DLPFC (right BA 46) (Greene et al., 2008b). In a second study (Greene et al., 2008d), we compared utilitarian and non-utilitarian/deontological moral disapproval. The footbridge dilemma typically elicits deontological disapproval (“It’s wrong to kill the one, despite the greater good,”). One can generate utilitarian disapproval using dilemmas like the reverse switch case, in which one can divert the trolley onto five people in order to save one person (an action that makes no utilitarian sense). Consistent with the dual-process theory, we found that utilitarian disapproval, as compared to deontological disapproval, was associated with greater activity in the same region of DLPFC as above. It is worth noting that the region of DLPFC associated with utilitarian judgment in these studies (BA 46) is posterior to that associated with utilitarian judgment in response to high-conflict personal moral dilemmas (BA 10) (Greene et al., 2004). All utilitarian judgments appear to require utilitarian reasoning, but additional cognitive control is only required in the face of countervailing emotional responses. Thus, it is possible that BA 46 is engaged during utilitarian moral reasoning, while BA 10 is engaged in the more extended cognitive processing elicited by high-conflict personal dilemmas. Finally, as noted above, Glenn and colleagues (2008) found that individuals with high psychopathy scores exhibited reduced amygdala activity during the contemplation of personal moral dilemmas, thus providing further evidence for the connection between emotion and deontological impulses (which are reliably
generated by personal moral dilemmas). They also found that individuals with high scores on the interpersonal factor of the Psychopathy Checklist (which involves manipulation, conning, superficiality, and deceitfulness) (Hare, 2003) exhibited decreased activation in the mPFC/PCC/TPJ network. (Note, however, that the psychopaths did not exhibit abnormal moral judgment behavior, complicating this interpretation.)

In sum, the dual-process theory of moral judgment, which emphasizes both emotional intuition and controlled cognition, is supported by multiple fMRI studies using different behavioral paradigms, multiple behavioral studies of neurological patients, and a variety of behavioral studies using both experimental manipulations and individual difference measures. (For an alternative perspective see Moll & de Oliveira-Souza (2007). For my reply, see Greene (2007a).)

The mental states of moral agents

As Oliver Wendell Holmes Jr. famously observed, even a dog knows the difference between being tripped over and being kicked. Holmes’ comment highlights the importance of information concerning the mental states of moral agents and, more specifically, the distinction between intentional and accidental harm.
Berthoz and colleagues (2006) identified several brain regions that exhibit increased activity in response to intentional (vs. accidental) moral transgressions, including the amygdala, the precuneus, the ACC, and the DLPFC. These results suggest a kind of dual-process response to intentional harms. That may be correct, but a more recent set of studies complicates this picture. Young and colleagues (2007) compared neural responses to intended harms, accidental harms, failed attempted harms, and ordinary harmless actions (a 2 x 2 design crossing mental state information (agent did / did not anticipate harm) and outcome information (harm did / did not result)). They found that the mPFC, PCC, and TPJ, all regions associated with theory of mind (Saxe et al., 2004), were not only sensitive to belief (i.e. anticipation) information, but were also sensitive to the interaction between belief and outcome information. More specifically, the right TPJ was particularly sensitive to attempted harm, consistent with the behavioral finding that attempted harm is readily condemned, while accidental harm is not so readily excused. (See also Cushman et al. (2006).) Interestingly, Young and colleagues found that judgments in response to accidental harm (as compared to intentional harm) were associated with increased activity in the ACC and DLPFC, regions associated respectively with conflict and control in the context of moral judgment (Greene et al., 2004). Young and colleagues argue that this is due to a conflict between an outcome-based response (the person caused harm) and one based on mental states (it was an accident). Thus, we see here increased activity suggestive of cognitive conflict
and control in response to accidental harms, as opposed to intentional harms (Berthoz et al., 2006).

Further studies by Young & Saxe have examined the roles of various neural regions in processing mental state information in the context of moral judgment. They have found that the mPFC is sensitive to the valence of the agent’s belief, while the TPJ and precuneus appear to be critical for the encoding and subsequent integration of belief information in moral judgment (Young and Saxe, 2008). A third study (Young and Saxe, in press) suggests that the right TPJ, PCC and mPFC are involved in the generation of spontaneous mental state attributions. Finally, they found, as predicted, that disrupting activity in the right TPJ using TMS produces a more child-like pattern of moral judgment (Piaget, 1965) based more on outcomes and less on mental state information (Young et al., 2008).

While most humans (and perhaps some canines) are explicitly aware of the distinction between intended and accidental harm, people’s judgments are also sensitive to a more subtle distinction between harms that are intended as a means to an end and harms that are merely foreseen as side-effects (Aquinas, unknown/2006). The means/side-effect distinction is, in fact, a key distinction that distinguishes the footbridge dilemma (in which a person is used as a trolley-stopper) from the switch dilemma (in which the a person is killed as “collateral damage”) (Foot, 1978; Thomson, 1985; Mikhail, 2000; Cushman et al., 2006; Moore et al., 2008). (Recent research suggests that the means/side-effect
distinction interacts with factors related to “personalness” in generating the effects that give rise to the Trolley Problem (Greene et al., 2008c). Schaich-Borg and colleagues (Schaich Borg et al., 2006) found that the anterior STS and VMPFC exhibit increased activity in response to dilemmas in which the harm is an intended means, as opposed to a foreseen side-effect. They also found increased DLPFC activity associated with harms caused through action, as opposed to inaction, consistent with the finding that people appear to have conscious access to the action/inaction distinction (Cushman et al., 2006).

While the studies described above highlight the importance of mental state representation in moral judgment, a study of moral judgment in autistic children indicates some basic moral judgments do not depend on theory of mind abilities (Leslie et al., 2006).

**Neuroeconomics**

Morality, broadly construed, may be viewed as a set of psychological adaptations that allow individuals to reap the benefits of cooperation (Darwin, 1871/2004). In economics, the most widely-used experimental paradigm for studying cooperation is the prisoner’s dilemma (Axelrod and Hamilton, 1981), in which two individuals maximize their total payoffs by cooperating, but maximize their individual payoffs by not cooperating (“defecting”). Rilling and colleagues found that brain regions associated with reward (nucleus accumbens,
caudate nucleus, VMPFC and OFC, and rostral ACC) were associated with cooperation, indicating that cooperative behavior is supported by general-purpose reward circuitry. A more recent study (Moll et al., 2006) in which people made charitable donations from inside the scanner teaches a similar lesson. Decisions to make costly donations were associated with increased activity in reward-related brain regions overlapping with those identified by Rilling and colleagues. This study also found a remarkably high correlation ($r = .87$) between self-reported engagement in voluntary activities and the level of activation in the mPFC during costly donation.

Several neuroeconomic experiments have used the Ultimatum Game (UG) (Guth et al., 1982) to examine neural responses to unfairness. In the UG, one player (the proposer) makes a proposal about how to divide a fixed sum of money between herself and the other player (the responder). The responder may either accept the proposal, in which case the money is divided as proposed, or reject it, leaving both players with nothing. Responders typically reject offers substantially below half of the total as unfair. Sanfey and colleagues (2003) found that responders responded to such unfair offers with increased activity in the insula, which is associated with autonomic arousal (Critchley et al., 2000) and negative emotion (Calder et al., 2001). The level of insula activity scaled with the magnitude of the unfairness, responded more to human- vs. computer-generated proposals, and was associated with higher levels of rejection. Unfair offers also elicited increased activity in the right DLPFC, which was interpreted as involved
in inhibiting the negative emotional response to unfairness. A more recent study (Knoch et al., 2006), however, challenges this interpretation, finding that disrupting activity in the right DLPFC generated fewer rejections of unfair offers. These results suggest that the right DLPFC is involved in inhibiting the appetitive desire for more money, rather than the punitive response to unfair treatment. Koenigs et al. (2007) found that patients with VMPFC damage exhibited the opposite behavioral pattern, suggesting that the VMPFC plays a critical role in regulating the emotional response that drives individuals to respond punitively to unfair treatment. (Increased rejection rates can also be generated by decreasing seratonin levels through tryptophan-depletion (Crockett et al., 2008).) A more recent fMRI study of the UG (Tabibnia et al., 2008) found that increased activity in the ventrolateral PFC is correlated with increased acceptance of unfair offers, suggesting that this region may play the role originally attributed to the right DLPFC.

Other neuroeconomic studies have focused on how individuals track and respond to the moral status of others. Singer and colleagues (2004) examined neural responses to faces of people who had played either fairly (i.e. cooperatively) or unfairly in a sequential prisoner’s dilemma game. Surprisingly, they found that faces of fair players, but not unfair players, elicited increased activity in the insula and the amygdala, regions widely, but not exclusively, associated with negative affect (Adolphs, 1999). In a second study, Singer and colleagues (2006) examined the interaction between (un)fair behavior
and empathy. Both males and females exhibited signs of pain-empathy (increased activity in the fronto-insular cortex and ACC) when observing fair players receive a painful shock, but this effect was significantly reduced in males when the players receiving the shock had played unfairly. Males, moreover, exhibited increased reward-related activity in the nucleus accumbens (correlated with self-reported desire for revenge) when observing unfair players get shocked. In a similar vein, de Quervain and colleagues (2004) observed that reward-related activity in the caudate nucleus was associated with willingness to punish individuals who betrayed the subject’s trust in a trust game. (A trust game is essentially a sequential prisoner’s dilemma game in which cooperators must trust one another to continue cooperation.) A study by Delgado and colleagues (2005) examined the effect of reputation on moral-economic interaction. They had subjects play a trust game with fictional individuals who were characterized as good, bad, or neutral based on their personal histories. Their reputations affected subjects’ willingness to trust them and modulated the level of activity in the caudate nucleus, partially overriding the effect of feedback during the game. King-Casas and colleagues (2006) used a trust game to examine the temporal dynamics of trust-development. They found that reward-related signals in the dorsal striatum were associated with the intention to trust and were shifted earlier in time as trust developed over the course of the game, mirroring effects observed in non-social reinforcement learning (Schultz et al., 1997). Taking a more molecular approach to the understanding of trust, Kosfeld
and colleagues (2005) found that intranasal administration of oxytocin, a neuropeptide known for its role in social attachment and affiliation in non-human mammals (Insel and Young, 2001), increased trusting behavior.

Hsu and colleagues (2008) examined the neural bases of decisions concerning distributive justice, pitting deontological considerations for equality against utilitarian considerations in favor of maximizing aggregate benefits (“efficiency”). Their subjects allocated money to children in an orphanage, with some options favoring equality at the expense of efficiency and vice versa. Aversion to inequality was associated with increased activity in the insula, while activity in the putamen was positively correlated with the efficiency of the outcome. The caudate nucleus, in contrast, was sensitive to both factors, reflecting the subjective utility of the option. While at odds with the relatively simple dual-process theory presented above, these results are consistent with the Humean (1739/1978) conjecture (Greene et al., 2004; Greene, 2007b) that both deontological and utilitarian considerations ultimately have affective bases, despite the latter’s greater dependence on controlled cognitive processing.

Conclusion

People often speak of a “moral sense” or a “moral faculty” (Hauser, 2006), but there is no single system within the human brain that answers to this description.
Rather, moral judgment emerges from a complex interaction among multiple neural systems whose functions are typically not (and maybe not ever be) specific to moral judgment (Greene and Haidt, 2002). The bulk of the research discussed above rightly emphasizes the role of emotion, in all of its functional and anatomical variety. At the same time, it is clear that controlled cognitive processing plays an important role in moral judgment, particularly in supporting judgments that run counter to prepotent emotional responses.

Three positive trends emerge from the foregoing discussion: First, we have seen a shift away from purely stimulus-based studies in favor of studies that associate patterns of neural activity with behavior. Second, and relatedly, we have seen an increased reliance on behavioral data, both in neuroscientific research and complementary behavioral studies. Third we have developed more ecologically valid paradigms involving real decisions, while recognizing that more stylized, hypothetical decisions can, like the geneticist’s fruit fly, teach us valuable lessons. With regard to this issue, it is worth noting that in modern democracies our most important decisions are made indirectly by voters whose individual choices have little bearing on outcomes, and are thus effectively hypothetical.

Our current neuroscientific understanding of moral judgment is rather crude, conceptualized at the level of gross anatomical brain regions and psychological processes familiar from introspection. But, for all our ignorance, the physical basis of moral judgment is no longer a complete mystery. We’ve not
only identified brain regions that are “involved” in moral judgment, but have begun to carve the moral brain at its functional joints.

Notes

1. We defined “personal” moral dilemmas/harms as those involving actions that are (a) likely to cause serious bodily harm, (b) to a particular person, where (c) this harm does not result from deflecting an existing threat onto a different party (Greene et al., 2001). The first two criteria respectively exclude minor harms and harms to indeterminate “statistical” individuals. The third criterion aims to capture a sense of “agency,” distinguishing between harms that are “authored” rather than merely “edited” by the agent in question. Recent research suggests that the dilemmas originally classified as “personal” and “impersonal” may be fruitfully classified in other ways (Mikhail, 2000; Royzman and Baron, 2002; Cushman et al., 2006; Waldmann and Dieterich, 2007; Moore et al., 2008).

Acknowledgments: Many thanks to Joe Paxton and Liane Young for their helpful comments.


Rev Neurosci 2:352-363.

Guilford 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.

Cohen JD, Perlstein WM, Braver TS, Nystrom LE, Noll DC, Jonides J, Smith EE (1997) 
Temporal dynamics of brain activation during a working memory task. Nature 
386:604-608.

York: Cambridge University Press.

generation and representation of galvanic skin conductance responses: a 


Cushman F, Young L, Hauser M (2006) The role of conscious reasoning and intuition in 
moral judgment: testing three principles of harm. Psychol Sci 17:1082-1089.


Kant I (1785/1959) Foundation of the metaphysics of morals. Indianapolis: Bobbs-Merrill.


43


Fig 1

a. Medial prefrontal cortex (BA 9/10)
   - Posterior cingulate/precuneus (BA 31)
   - Superior temporal sulcus/temperoparietal junction (BA 39)

b. Inferior frontal gyrus (BA 44)
   - Middle frontal gyrus (BA 46)
   - Posterior cingulate cortex (BA 23/31)
   - Inferior parietal lobe (BA 40)

c. Precuneus (BA 7/31)
   - Anterior cingulate cortex (BA 32)
   - Posterior cingulate cortex (BA 23/31)

d. Superior/middle frontal gyrus (BA 10)
**A Dual-Process Theory of Moral Judgment**

- **a. utilitarian moral reasoning**
  *dorsolateral PFC, BA 46*
  *Inferior parietal lobe, BA 7/40*
  (Greene et al., 2001, 2004, 2008a, 2008b; Bartels, 2008; Moore, 2008; Hardman, 2008)

- **b. utilitarian cognitive control**
  *dorsolateral PFC, BA 10*
  (Greene et al., 2004, 2008c)

- **c. conflict monitor**
  *anterior cingulate cortex*
  (Greene et al., 2004, 2008)

- **d. intuitive appraisal**
  *Amygdala?, superior temporal sulcus/TPJ?*
  (Greene et al., 2001, 2004; Glenn et al., 2008)

- **e. deontological/non-utilitarian intuitive emotional response**
  *medial PFC, BA 9/10; posterior cingulate cortex/posterior STS/TPJ ("default" regions)*

**Fig 2**
Fig 3