CHAPTER

47 Moral Thinking

Liane Young

Abstract

This chapter presents several current models of moral thinking, with a focus on the cognitive processes that support people's moral judgments and justifications. These models are not mutually exclusive; rather, based on recent evidence from psychology and neuroscience, they posit different cognitive processes as the primary source of moral thinking. This chapter therefore does not quantify the evidence for one model versus another but instead reviews evidence for each model separately. These models, discussed in turn, emphasize the role of conscious principled reasoning, emotional processing, theory of mind, and a domain-specific "moral faculty" that integrates information from other cognitive systems to compute specifically moral judgments.

Key Words: moral judgment, justification, reason, emotion, theory of mind, moral faculty

The topic of morality has been of interest to philosophers and indeed ordinary people long before cognitive psychologists and neuroscientists took up their tools to investigate the moral mind and brain. Moral thinking as a topic for empirical science is both rewarding and challenging precisely because people—scientists or not—think about moral thinking. What sorts of behaviors are morally right or wrong? How do we make these judgments? Are there right or wrong ways to go about this? Many people find these questions easy to answer simply by introspecting on their own experience, rather than relying on experiments, either scientific experiments, or, in the case of philosophy, thought experiments. In contrast to vision, language, or motor control, morality appears accessible to everyone—not only do we engage in moral thinking, but we also have a few thoughts on how we do so. For some people, moral thinking feels a lot like thinking: It involves thoughtfully considering pros and cons and rationally reflecting on moral principles. For others, moral thinking reduces to moral feeling: Some things feel right, other things feel wrong, and our emotions help us track important moral distinctions. Some people recognize culture and education as the primary sources of moral thinking; others appeal to an innate sense of right and wrong, an unconscious moral code. Scientists are people too, and so, not surprisingly, the last decade has seen a frenzy of empirical activity as we begin to put many of these intuitions to the test.

In this chapter, I will present several current models of moral thinking, with a particular focus on how we think about and make moral judgments and justify them, rather than how we actually behave because of or in spite of these judgments. I will discuss evidence from cognitive psychology and neuroscience for each of these models, named here for their primary focus: (1) Reason, (2) Emotion, (3) Theory of Mind (i.e., the processing of mental states such as beliefs and intentions), and (4) Moral Faculty. Importantly, these models are not mutually incompatible. For instance, not all Emotion models will deny roles for Reason or Theory of Mind; Emotion models simply emphasize Emotion as the dominant process in moral judgment. Or, in

the case of Theory of Mind, both domain-specific (e.g., specific to Theory of Mind) and domain-general processes (e.g., Reason) may contribute to the influence of mental state factors on moral judgment; however, here, we will focus on evidence for the domain-specific contributions. In the following four sections, I will therefore present evidence for the roles of reason, emotion, theory of mind, and a moral faculty in moral thinking.

Reason: Moral Thinking Is "Thinking"

On a reason model, moral thinking is dominated by "thinking"—of the conscious, controlled sort. In other words, most of the time, for most of moral thinking, people consult explicit moral principles or theories, engage in conscious reasoning, and in general behave as rational agents. People therefore make moral decisions that they would endorse upon reflection, given the important role of reflection in the first place. Developmental psychologists such as Piaget and Kohlberg supported such reason-based models of moral judgment and, as a consequence, identified participants' ability to articulate justifications for their moral judgments as the primary indication of moral maturity (Kohlberg, 1981; Piaget, 1932/1965). The ability to engage in conscious, principled moral reasoning was supposed to track with stages of moral development, since moral judgment was supposed to reflect directly the reasoning that led to it, and not any funny business operating under our conscious radar.

Contemporary moral psychology arose largely in resistance to rationalist or reason-based models of moral judgment (Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, Koller, & Dias, 1993). As a result, much of the evidence that follows for moral thinking as "thinking" is indirect and, in fact, falls out of results primarily seen as supporting a different conclusion, that is, moral thinking is moral feeling. That most of the evidence for moral "thinking" is indirect is also notable. At first, it might appear easy to generate evidence for moral "thinking"—after all, if moral thinking is conscious and controlled, then participants should be able to detect and report engaging in moral thinking. However, the question at hand is not whether people ever consult explicit moral principles or theories, but rather whether conscious moral reasoning is the causal source of moral judgment. This question turns out to be more difficult to address—people may report having made particular moral judgments for particular moral reasons; however, based on this report alone, the experimenter cannot know whether participants generated those reasons post hoc, after the fact, to rationalize or justify their judgments, or whether those reasons did causally determine those judgments. Most of contemporary moral psychology has emphasized the surprising absence of "thinking" in moral thinking, as we will see in the next section. This section, however, serves to show that thinking may play a greater role in moral judgment than has been recently thought—and in surprising contexts—contexts used to show that feeling, not thinking, dominates moral psychology.

The reason model makes two basic predictions. The first is that moral judgments, moral justifications, and, critically, their correspondence are subject to the influence of demographic variables such as education and culture, which partly determine access to, as well as reliance on, reason. In the first of an important line of studies, Jonathan Haidt and his colleagues investigated moral judgment and justification in participants of low and high socioeconomic status (SES) from Brazil and Philadelphia (Haidt et al., 1993). Participants were asked to judge not only harmful actions but also harmless actions that nevertheless violated norms of moral purity, such as consensual incest, eating the family dog upon its demise, and other disgusting but victimless transgressions. Contrary to the reason model, the upshot of this and related work is that many participants are unwilling to endorse taboo violations that elicit strong emotional responses even when they are unable to articulate reasons for their judgments—for instance, why incest is morally wrong even in the absence of any physical or psychological harm (Haidt, 2001).

This particular study, however, allows for a closer look, as its title suggests, at affect, culture, and morality (Haidt et al., 1993). Indeed, high SES participants from Philadelphia (described in the study as college students at elite universities), as compared to low SES participants from Brazil, were more likely to endorse harmless actions they found disgusting. By contrast, low SES participants from Brazil continued to judge these taboo violations as morally forbidden even when they were unable to justify their judgments, thereby revealing poor correspondence between their judgments and justifications. These demographic differences in judgments and justifications were also observed to be more pronounced in adults than children. Over time, then, differences in education and culture may lead to differences in the reliance on conscious principled reasoning for moral judgment and perhaps

the resistance to moral judgments that appear not to be based on reason. This study therefore suggests that reasoning abilities, as modulated by culture and education, impact moral judgment even in the presence of strong emotions such as disgust.

A related body of research reveals a dissociation between implicit and explicit attitudes toward race and sexual orientation (Banaji, 2001; Inbar, Pizarro, Knobe, & Bloom, 2009). Very liberal college students from Berkeley, for example, appear capable of overriding their negative emotionally mediated implicit attitudes toward gay and interracial sex in order to explicitly endorse gay and interracial sex across a number of behavioral measures (Inbar, Pizarro, Knobe, et al., 2009). Of course, implicit attitudes may constitute moral judgment in some sense, too. However, to the extent that people's explicit moral attitudes, determined in part by culture and education, drive moral judgment and behavior, these findings are consistent with the rational correction of implicit emotionally mediated attitudes.

The second prediction is related to the first: People make moral judgments based on factors that they endorse as morally relevant (e.g., the distinction between harming via action versus omission), and, correspondingly, people are willing to reject moral judgments made on the basis of factors they regard as morally irrelevant (e.g., the distinction between harming via personal contact versus no contact). Research from Fiery Cushman and his colleagues has provided evidence of conscious reasoning from moral principles to moral judgments (Cushman et al., 2006). One such principle, *it is morally worse to harm by action than omission*, was articulated by participants when required to justify their judgments of particular scenarios, in particular, their response that killing a person in one scenario was worse than letting a person die in another scenario. Because participants were able to articulate this general principle, it is at least possible that they consciously reasoned from this principle to their moral judgments; however, the alternative is that they reconstructed this principle post hoc when required to justify their prior judgments. Importantly, against this alternative, participants who were able to articulate this principle in their justifications showed significantly greater use of this principle in their judgments. Meanwhile, when participants in the same study discovered that their moral judgments were governed by a morally dubious principle (e.g., it is morally worse to harm via physical contact than no contact), they disavowed both the principle and their judgments that were based on the principle. Together, these findings indicate an impact of conscious principled reasoning on moral judgments.

Behavioral evidence from Tania Lombrozo also suggests a relationship between general moral commitments and moral judgments of particular scenarios (Lombrozo, 2009). Participants who explicitly endorsed consequentialist moral theories, that is, theories focused on the moral significance of consequences (e.g., the greatest good for the greatest number) were more likely to ignore nonconsequentialist distinctions between scenarios (e.g., physical contact versus no contact) when the scenarios were presented side by side. In fact, as will be discussed in more depth in the next section, an important body of behavioral and neural evidence suggests a direct correspondence between conscious principled reasoning and consequentialist moral judgments (Greene et al., 2001; Koenigs et al., 2007; Valdesolo & DeSteno, 2006). Extensive work by Josh Greene and his colleagues suggests a correlation between consequentialist moral judgments and activity in brain regions for abstract reasoning, such as the dorsolateral prefrontal cortex (DLPFC), as well as slower consequentialist moral judgments under cognitive load (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004).

In all, these results reveal a substantial role for reason in moral judgment—even in the very cases thought to exemplify the emotion model. Reasoning abilities, determined in part by culture and education, may allow us in some instances to override our initial moral attitudes. As rational moral agents, we may be able to make judgments based on principles and theories that we explicitly endorse and invoke when justifying our judgments. As we will see in the next section, though, not all moral judgments arise from conscious, principled reasoning. Much of moral thinking may be emotionally mediated and immune to conscious correction.

Emotion: Moral Thinking Is Moral Feeling

The emergence of the Emotion Model accompanied the birth of contemporary moral psychology (Haidt et al., 1993). On this model, most of moral thinking is moral feeling—judgments are made not via conscious, principled reasoning primarily but via emotional responses. These emotions include those, like disgust, that drive Haidt's participants to condemn disgusting but harmless actions (e.g., eating the family dog), as we will discuss next, as well as prosocial emotions, like empathy, disrupted

in certain patient populations (e.g., psychopathy, frontotemporal dementia, ventromedial prefrontal damage), as we will discuss later in this section.

Emotion models do not necessarily deny a role for other cognitive processes, as discussed in the prior section ("Reason) and subsequent section ("Theory of Mind"). Haidt, for example, recognizes a limited role for conscious reasoning—in social contexts and motivated moral reasoning. Nevertheless, on Haidt's model, emotions such as disgust dominate, as evident in the title of his seminal paper: "The Emotional Dog and Its Rational Tail" (Haidt, 2001). Conscious reasoning plays a more prominent role in Greene's dual-process model, as discussed later in this section (Greene et al., 2004). Greene's model highlights the competitive interaction between consciously reasoned responses (e.g., consequentialist responses) and emotional responses, including empathy. This section discusses the behavioral, neuroimaging, and neuropsychological evidence for the contribution of emotional processes to moral judgments.

Disgust, in particular, appears to be a key candidate emotion for certain moral judgments. On the one hand, certain actions may be perceived as immoral but not disgusting (e.g., tax fraud) and other actions as purely disgusting but not immoral (e.g., drinking urine). On the other hand, abundant research indicates a complex and often causal relationship between disgust and morality. Recent research suggests, for example, that individual differences in disgust responses may drive moral judgments: Political conservatives are more "disgust sensitive" (Inbar, Pizarro, & Bloom, 2009), a trait that can play a causal role in moral attitudes toward homosexuality (Inbar, Pizarro, Knobe, et al., 2009). In a more direct test of the causal link between disgust and moral judgment, participants were hypnotized to experience disgust at an arbitrary word (e.g., "often") and consequently delivered harsh moral judgments even when this word described benign behaviors (e.g., the student often chose popular topics for discussion) (Wheatley & Haidt, 2005). A related study investigated the impact of disgust induction on moral judgment (i.e., via a disgusting smell, a disgusting testing room, a memory of a physically disgusting experience, and a disgusting video) (Schnall, Haidt, Clore, & Jordan, 2008). After the disgust induction, subjects made harsher moral judgments, particularly if they were sensitive to their own bodily state. Finally, in a surprising study of moral behavior, participants were found to be more likely to engage in physical cleansing after behaving immorally and, conversely, to engage in immoral behaviors after physical cleansing (Zhong & Liljenquist, 2006). These behavioral findings converge on the notion that our emotional responses, and especially disgust, can dramatically shape our moral thinking.

Studies using functional magnetic resonance imaging (fMRI) have also revealed an important association between activity in brain regions implicated in disgust, including the insula, and moral judgments of purity violations (e.g., incest) (Schaich Borg, Lieberman, & Kiehl, 2008) and even unfair and harmful actions (Hsu, Anen, & Quartz, 2008; Moll et al., 2005). More generally, since its inception, contemporary moral psychology has seen a continuous stream of neuroimaging and neuropsychological research focused on the role of emotion in moral judgment (Young & Koenigs, 2007).

Much of this research has focused on the role of brain regions involved in empathy and emotional responsiveness, in particular, the ventral and medial portions of prefrontal cortex, referred to as ventromedial prefrontal cortex (VMPC). The VMPC includes the medial portions of orbitofrontal cortex (Brodmann areas 11 and 12) and the medial prefrontal cortex from the ventral surface to around the level of the genu of the corpus callosum (Brodmann area 25 and portions of Brodmann areas 10 and 32). The VMPC projects to limbic, hypothalamic, and brainstem regions that execute visceral and autonomic components of emotional responses (Ongur & Price, 2000); neurons within the VMPC encode the emotional value of sensory stimuli (Rolls, 2000). Damage to VMPC results in striking impairments in emotional function, including generally blunted affect, diminished empathy, emotional lability, and poorly regulated anger and frustration (Anderson, Barrash, Bechara, & Tranel, 2006; Barrash, Tranel, & Anderson, 2000). Activation in the VMPC is therefore taken as evidence of emotional processing, as in many of the fMRI studies described next.

Early fMRI studies of moral thinking were designed to isolate whatever cognitive processes or neural substrates might be specific to morality. With domain specificity in mind, these studies relied on paradigms contrasting neural responses to moral stimuli versus nonmoral stimuli. In an early study, subjects viewed emotionally salient scenes with moral content (e.g., physical assaults, war scenes) versus nonmoral content (e.g., body lesions, dangerous animals; Moll, de Oliveira-Souza, Bramati, & Grafman, 2002). Regions of VMPC, in this case the right medial orbitofrontal cortex and medial

frontal gyrus (Brodmann areas 10 and 11), were selectively recruited when participants passively viewed the moral scenes. Broadly similar activation patterns in the VMPC (lower medial Brodmann area 10) were observed when moral and nonmoral stimuli were matched for social content (e.g., number of people depicted in the scenes) and when subjects down-regulated their own emotional responses to the moral stimuli (Harenski & Hamaan, 2006). Another series of studies targeted the relationship between emotion and explicit moral judgment, replacing moral scenes with "moral statements," simple descriptions of morally salient behavior. VMPC activation (left medial orbitofrontal cortex) was selectively enhanced during the processing of emotionally salient moral statements (e.g., "He shot the victim to death") versus socially and emotionally salient nonmoral statements (e.g., "He licked the dirty toilet") (Moll, de Oliveira-Souza, Eslinger, et al., 2002). Even in the absence of explicit emotional content, moral statements describing morally inappropriate or appropriate actions (e.g., "A steals a car"/"A admires a car") elicited enhanced VMPC activation (medial Brodmann area 10) compared to nonmoral statements that were either semantically appropriate or inappropriate (e.g., "A takes a walk"/"A waits a walk") (Heekeren, Wartenburger, Schmidt, Schwintowski, & Villringer, 2003). Finally, in a similar vein, VMPC activation was observed for silent "right" or "wrong" judgments of simple statements with moral content (e.g., "We break the law when necessary") versus nonmoral content (e.g., "Stones are made of water") (Moll, Eslinger, & de Oliveira-Souza, 2001). Across these fMRI studies, emotional brain regions, in the VMPC, were recruited for moral thinking, in particular, the processing of moral scenes and statements versus nonmoral scenes and statements controlled for social and emotional content. These early studies set the stage for addressing additional questions about emotion and moral judgment in more detail. Do emotions support the processing of complex moral stimuli such as moral dilemmas? Do emotions systematically drive specific moral judgments?

Greene and his colleagues were the first to investigate whether emotion-related areas, such as the VMPC, support moral judgment in the context of moral dilemmas and, importantly, whether neural activity tracks with different moral content, within the moral domain. An early topic of investigation was the difference between moral scenarios that were "personal" or more emotionally salient and moral scenarios that were "impersonal" or less emotionally salient (Greene et al., 2001). For example, a trolley is headed for five people, and participants can save them by sacrificing the life of one person instead. In the "impersonal" scenario, participants can choose to turn the trolley away from the five people onto a side track where one person where will be hit instead. In the "personal" scenario, participants can choose to push a large stranger off a footbridge onto the tracks below, where his body will stop the trolley from hitting the five, though he, of course, will be hit. Personal moral scenarios selectively recruited VMPC (medial Brodmann area 10).

Greene and colleagues took this finding further when they investigated whether the observed activation patterns track not only emotionally salient scenarios but also emotionally mediated moral judgments. In particular, does emotional engagement track nonconsequentialist moral judgments—judgments based on factors other than consequences (e.g., intention, physical contact)? In Greene's experiments, nonconsequentialist judgments consisted of rejecting harmful actions that maximized good consequences (e.g., killing one to save five), while consequentialist judgments consisted of endorsing such harmful actions (Greene et al., 2004). Participants therefore made judgments for a series of scenarios. For example: 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? For this scenario, the nonconsequentialist judgment (e.g., "Don't smother the baby") is to reject the harmful action, even though it would maximize the greater good. On Greene's model, this judgment is rooted in an automatic emotional aversion to the harmful act. In other words, participants' prepotent response is an emotional response to the harm, leading participants to reject it. By contrast, consequentialist reasoning requires participants to stifle their emotional response (e.g., "The baby will die no matter what, so smother the baby to save everyone else"), leading to the consciously reasoned and emotionally incongruent judgment. Consistent with Greene's

dual-process model, brain regions associated with cognitive conflict and abstract reasoning, such as anterior cingulate and dorsolateral prefrontal cortex, were selectively recruited for consequentialist judgments. Subjects appeared able to override their emotional aversion to the harm and engage in consequentialist reasoning.

The growing body of neuroimaging work suggests emotional engagement during moral judgment and, in particular, nonconsequentialist moral judgments. Neuroimaging methods, however, are currently limited in that they reveal only correlations between neural activity and cognitive processes rather than causally necessary connections. One way to determine whether emotional processing plays a causally necessary role in moral judgment is to test individuals with selective deficits in emotional processing. As we will see, neuropsychological studies suggest a causal connection between prosocial emotions and social cognition. Emotional dysfunction often does lead to deficits in moral judgment, reasoning, and behavior.

An early study investigated moral reasoning in two adult individuals with early-onset VMPC lesions (Anderson, Bechara, Damasio, Tranel, & Damasio, 1999). According to a traditional characterization of moral development (Kohlberg, 1981), both individuals exhibited a "preconventional" stage of moral reasoning. More specifically, both early-onset VMPC lesion patients provided immature moral justifications, engaging in moral reasoning from an egocentric perspective of punishment-avoidance (e.g., reasoning that stealing medicine for a loved one is morally wrong because one might get caught). This finding was especially striking given prior research revealing normal moral reasoning in patients with adult-onset VMPC damage (Saver & Damasio, 1991). The moral reasoning deficit documented in the early-onset cases suggests that areas of prefrontal cortex support the original acquisition of normal moral reasoning abilities (Anderson et al., 1999). A caveat, however, is that these studies target participants' justifications rather than judgments, licensing only limited conclusions about the role of the VMPC in moral thinking.

Investigations of adult and developmental psychopaths have associated emotional impairment with defects in moral behavior and judgment. Psychopathy is typically associated with pronounced emotional impairment, for instance, considerably reduced empathy and guilt, and pronounced behavioral disturbance, for instance, criminal and frequently violent behavior (Hare, 1991). Reports of deficits in prosocial emotions (e.g., empathy) and behavior motivated a pair of studies on moral judgment in psychopathy (Blair, 1995, 1997). James Blair found that both adult and developmental psychopaths were unable to distinguish between unambiguous moral transgressions (e.g., hitting someone) and unambiguous conventional transgressions (e.g., talking out of turn) along the dimensions of permissibility, seriousness, and authority contingence—a distinction that even young children are able to make (Turiel, 1983). Recent work suggests that individuals with psychopathic tendencies may show normal moral judgments in limited contexts but simply lack motivation to behave prosocially in accordance with their judgments (Cima, Tonnaer, & Hauser, 2010).

The first direct investigation of moral judgment (as opposed to behavior or justification) in brain-damaged populations was a study of patients with frontotemporal dementia (FTD; Mendez, Chen, Shapira, & Miller, 2005). FTD involves deterioration of prefrontal and anterior temporal brain areas. FTD patients therefore exhibit blunted emotion and diminished regard for others early in the course of the disease. Similar to psychopathy, FTD is marked by behavioral changes, including transgressive behavior, that is, stealing, physical assault, and inappropriate sexual advances (Mendez, Chen, et al., 2005). In light of the deficits in prosocial emotions and behavior associated with FTD, Mendez and colleagues (Mendez, Anderson, & Shapira, 2005) investigated FTD patients' moral judgments of personal and impersonal moral scenarios (Greene et al., 2001). Again, most healthy participants advocate turning a trolley away from five people and onto one person, in the impersonal scenario, but not pushing a stranger off a footbridge so that his body will stop a trolley from hitting five people, in the personal scenario (Cushman et al., 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2002) However, most FTD patients endorsed the harmful action for both the impersonal and personal scenarios. Social psychologists have observed similar patterns of judgment after reducing negative affect by exposing subjects to Chris Farley's comedic *Saturday Night Live* skits (Valdesolo & DeSteno, 2006). These results suggest that due to the deterioration of emotional processing mediated by the VMPC, the FTD patients did not fully experience the emotional salience of the personal harm (e.g., pushing the stranger). However, since neurodegeneration in FTD affects multiple prefrontal

and temporal areas, precise conclusions about the impact of emotional processing subserved by the VMPC versus other cognitive functions cannot yet be drawn.

Investigating moral judgment in individuals with focal VMPC lesions represents the most direct approach to the relationship between emotional processing in the VMPC and moral judgment. Like FTD patients, VMPC lesion patients exhibit diminished empathy and blunted affect, but, importantly, unlike FTD patients, VMPC lesion patients retain broader intellectual function. VMPC patients can therefore be studied to characterize the specific role of emotion in moral judgment. One study 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, causally necessary for normal moral judgment (Koenigs et al., 2007). In this study, patients evaluated the same impersonal and personal moral scenarios described earlier. As in previous fMRI studies (Greene et al., 2001, 2004), many of the personal scenarios pit an emotionally aversive harm against the "greater good" (e.g., killing one to save many). Like the FTD patients, 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 if a greater number of people would benefit—the consequentialist judgment. A second lesion study confirmed this finding (Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007). Together, these studies suggest that emotional processing mediated by VMPC is crucial for moral judgment and in particular consequentialist moral judgment (Greene et al., 2004).

All of these studies, however, rely on moral scenarios describing intentional harms: Agents act with the belief and intent, stated or implied, that they will cause the harmful outcome that they, in fact, cause. It is thus unresolved whether the brain regions implicated in emotional processing, such as the VMPC, are involved in processing harmful outcomes or harmful intentions or both. Recent functional neuroimaging and neuropsychological evidence suggests that the VMPC supports the processing of harmful intentions (Young & Saxe, 2009b). In healthy adult participants, moral judgments of failed attempts to harm (harmful intention, neutral outcome) were significantly correlated with the average neural response in the VMPC. Individuals with a high VMPC response, and a stronger emotional response to the harmful intention, assigned more blame for failed attempts to harm, while individuals with a low VMPC response, and a weaker emotional response to the harmful intention, assigned less blame.

A follow-up study investigated moral judgments made by patients with adult-onset VMPC lesions, as in the study described earlier (Young, Bechara, et al., 2010). Consistent with the fMRI evidence, VMPC patients judged attempted harms as significantly more morally permissible, compared to control participants—and even compared to their own judgments of accidental harms. In fact, this pattern reflects a striking reversal of the normal pattern of moral judgments; in judging failed attempts to harm as more permissible than accidental harms, VMPC patients revealed an extreme "no harm, no foul" mentality. VMPC patients showed a selective deficit in moral judgment of attempted harms, including attempted murder, suggesting that although VMPC patients may be able to reason about the content of a belief or an intention, they are unable to trigger normal emotional responses to mental state content for moral judgment—in line with prior work showing deficits in their emotional processing of abstract versus concrete information (Bechara & Damasio, 2005). The finding that the VMPC is associated with processing intentions with high emotional content, that is, harmful intent, for moral judgment, is also consistent with the role of the VMPC in "affective" theory of mind or emotional empathy (Jenkins & Mitchell, 2009; Mitchell, Macrae, & Banaji, 2006; Shamay-Tsoory & Aharon-Peretz, 2007; Vollm et al., 2006). Prior evidence has suggested a specific role for the VMPC in processing affective aspects of another person's mental states (Jenkins & Mitchell, 2009; Mitchell et al., 2006; Shamay-Tsoory & Aharon-Peretz, 2007; Vollm et al., 2006). Therefore, damage to these processes for emotional empathy may lead to deficits in both moral judgment and prosocial behavior.

Research using behavioral methods, fMRI, and neuropsychology has illuminated the specific role of emotion in moral judgment. Manipulating emotions, by either enhancing or suppressing them, can systematically bias people's moral judgments; brain regions associated with emotional processing are recruited for moral judgment, and more for some kinds of moral judgments over others; patients with deficits in emotional processing show systematically abnormal moral cognition in judgment, justification, and behavior.

Theory of Mind: Moral Thinking Is Thinking About Thinking

A third cognitive model posits that, in addition to conscious reasoning and emotional processing, theory of mind is a key cognitive process for moral judgment, that is, how we reason about the mental states of moral agents, including their innocent and guilty intentions (Hart, 1968; Kamm, 2001; Mikhail, 2007). My colleagues and I have focused on the dominant role of mental states versus outcomes for moral judgment (Young, Cushman, Hauser, & Saxe, 2007). In our studies, participants typically read stories in which agents produced either a negative outcome (harm to another person) or a neutral outcome (no harm), based on the belief that they would cause the negative outcome ("negative" belief or intention) or the neutral outcome ("neutral" belief or intention). Participants then judge whether the action was morally permissible or forbidden, or how much moral blame the agent deserves.

For example, in one scenario, Grace and her coworker are taking a tour of a chemical factory. Grace stops to pour herself and her coworker some coffee. Nearby is a container of sugar. The container, however, has been mislabeled "toxic," so Grace thinks that the powder inside is toxic. She spoons some into her coworker's coffee and takes none for herself. Her coworker drinks the coffee, and nothing bad happens. In an alternative scenario, a container of poison sits near the coffee. The container, however, has been mislabeled "sugar," so Grace thinks the powder inside is sugar. She spoons some into her coworker's coffee. Her coworker drinks her coffee and ends up dead. Across all of our studies using scenarios like these, participants weighed the agent's belief and intent more heavily than the action's outcomes in their moral judgments (Young et al., 2007; Young, Nichols, & Saxe, 2010). A simple metric of this effect is that our participants almost universally judge an attempted harm (negative belief, neutral outcome) as more morally blameworthy and more morally forbidden than an accidental harm (neutral belief, negative outcome).

Cushman (2008) has pushed this line of work even further, directly comparing the roles of outcome, causation, beliefs, and desires for different kinds of moral judgments (e.g., person, permissibility, blame, and punishment) (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009). The agent's belief about whether his or her action would cause harm was the most important factor across the board, followed by the agent's desire to cause harm. Notably, though, judgments about how much to punish the agent relied relatively more on outcomes, as compared to judgments about the moral character of the agent or the moral permissibility of the action, which relied more on beliefs.

What's surprising is that mental state factors dominate even where external outcomes appear to drive moral judgments. For instance, agents who cause harmful outcomes by accident are still judged to be somewhat morally blameworthy in spite of their mental states. Consider again the scenario where Grace accidentally poisons her coworker because she mistakes poison for sugar. Though we mostly let Grace off the hook for her false belief and innocent intention, we still assign some moral blame. Recent research suggests that this negative moral judgment is based not simply on the harmful outcome of Grace's action but largely on participants' assessment of Grace's mental state (Young, Nichols, & Saxe, 2010). In particular, participants judge Grace's false belief as unjustified or unreasonable and therefore Grace as morally blameworthy. So, even when we assign blame for accidents, we may do so on the basis of mental state factors (e.g., negligence) and not simply on the basis of the harmful outcome.

For most healthy adults, mental states, including beliefs, intentions, and desires, carry more moral weight than external outcomes. In some cases, mental states overwhelm other morally relevant external factors, including external constraints, like whether the person could have acted otherwise (Woolfolk, Doris, & Darley, 2006). Woolfolk, Doris, and Darley (2006) presented subjects with variations of one basic story: Bill discovers that his wife Susan and his best friend Frank have been involved in a love affair. All three are flying home from a group vacation on the same airplane. In one variation of the story, their plane is hijacked by a gang of ruthless kidnappers, who surround the passengers with machine guns and order Bill to shoot Frank in the head; otherwise, they will shoot Bill, Frank, and the other passengers. Bill recognizes the opportunity to kill his wife's lover and get away with it. He wants to kill Frank and does so. In another variation: Bill forgives Frank and Susan and is horrified when the situation arises but complies with the kidnappers' demand to kill Frank. On average, observers rate Bill as more responsible for Frank's death, and the killing as more wrong, when Bill wanted to kill Frank, even though this desire played no role in causing the death, in either case.

While assigning moral blame for harmful desires and intentions appears easy and automatic (except in

case of VMPC damage), forgiving accidental harms appears to present more of a challenge. Among healthy adults, we have found evidence of substantial individual variability in moral blame assigned to protagonists for accidental harms (Young & Saxe, 2009b). In development, full forgiveness or exculpation for accidents does not emerge until approximately 7 years of age, surprisingly late in childhood. Meanwhile, 5-year-old children are capable of reasoning about false beliefs: In the paradigmatic "false-belief task," children predict that observers will look for a hidden object where they last saw the object, not in its true current location (Flavell, 1999; Wellman, Cross, & Watson, 2001). These same children, however, judge that if a false belief led an observer to unknowingly and accidentally cause harm to another person (e.g., mistake poison for sugar), the agent is just as bad as if he or she had caused the harm on purpose (Piaget, 1932/1965). The ability to integrate beliefs and intentions into moral judgments appears then to be a distinct developmental achievement (Young & Saxe, 2008). Consistent with this idea, high-functioning adults diagnosed with Asperger's syndrome, who also pass standard false-belief tasks, assign abnormally high levels of moral blame for accidental harms as well (Moran et al., 2011).

My colleagues and I have recently investigated the neural mechanisms that support moral judgments based on mental states such as beliefs and intentions. Our results suggest that specific brain regions support multiple distinct cognitive components of mental state reasoning for moral judgment: the initial encoding of the agent's belief, the use and integration of the belief (with outcome information) for moral judgment, as discussed earlier, spontaneous mental state inference when mental state information is not explicitly provided in the moral scenario, and even post hoc reasoning about beliefs and intentions to rationalize or justify moral judgments (Kliemann, Young, Scholz, & Saxe, 2008; Young et al., 2007; Young & Saxe, 2008, 2009a).

Building on prior research on neural substrates for theory of mind in nonmoral contexts (Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Kanwisher, 2003), our research suggests that the most selective brain region appears to be the right temporo-parietal junction (RTPJ). In one study, individual differences in moral judgments were correlated with individual differences in the RTPJ response (Young & Saxe, 2009b). Participants with a high RTPJ response, and a more robust mental state representation (e.g., false belief, innocent intention), assigned less blame to agents causing accidental harm. Participants with a low RTPJ response, and a weaker mental state representation, assigned more blame, like young children and our participants with Asperger's Syndrome. One source of developmental change in moral judgments may therefore be the maturation of specific brain regions for representing mental states such as beliefs—consistent with recent research suggesting the RTPJ may be late maturing (Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009).

The correlation observed here between the use of mental states for moral judgment and the neural response in a brain region dedicated to mental state reasoning suggests that individual differences in moral judgment are not due exclusively to individual differences in domain-general capacities for abstract reasoning or cognitive control, as discussed in the preceding sections. What determines blame or forgiveness is not just the ability to override a prepotent response to a salient harmful outcome (Greene et al., 2004). The conflict between mental state and outcome factors may account for part of the challenge of forgiveness. The neural data suggest that the strength of the mental state representation matters for how the conflict is resolved—and whether forgiveness or blame is offered.

Disrupting RTPJ activity also disrupts the use of mental state information for moral judgment. In a recent study, we produced a temporary "virtual lesion" in the RTPJ, using a neurophysiological technique known as transcranial magnetic stimulation (TMS) (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). TMS allows the induction of a current in the brain, using a magnetic field to pass the scalp and skull. After using fMRI to identify the RTPJ in each of our participants, and a nearby control region not implicated in mental state reasoning, we used offline and online TMS to modulate neural activity in two experiments. In both experiments, TMS to the RTPJ versus the control region made a significant and selective difference, reducing the impact of intentions and, as a direct result, increasing the impact of outcomes on moral judgments. For example, disrupting RTPJ activity led to more lenient judgments of failed attempts to harm, based on the neutral outcome, and not the harmful intent. Indeed, moral judgment depends critically on specific neural substrates for processing mental states like beliefs and intentions.

Together, these studies provide behavioral and neural evidence for theory of mind as a key cognitive process for moral judgment. Evaluating moral

agents and their actions requires an assessment of the agents' mental states. Specific neural substrates support such mental state assessments. Compromised mental state reasoning in the case of neurodevelopmental disorders (e.g., autism) or via TMS therefore leads to abnormal moral thinking.

Moral Faculty: Moral Thinking Is Uniquely Moral

In contrast to the models presented so far, moral faculty models focus on what, if anything, might be specific to the domain of morality—rather than the contribution of cognitive processes already known to operate in other domains (e.g., reasoning, emotion, theory of mind). Moral faculty models posit a uniquely moral faculty that takes multiple cognitive inputs and computes a uniquely moral judgment (Hauser, 2006; Mikhail, 2002, 2007). Interestingly, at the outset of contemporary moral psychology, the key questions for many focused directly on what was specific to the moral domain. Are moral judgments governed by specific moral rules and computations, or specific moral emotions? Are moral judgments supported by a specific neural substrate or network? An illustration of this domain-specific approach can be found in early fMRI studies of morality, described earlier in the section on "Emotion." Many of these studies were designed around comparing moral stimuli to nonmoral stimuli, while controlling for differences along other dimensions (e.g., emotional salience, social content). Before too long, though, moral psychology saw a subtle shift from these questions of domain specificity toward questions concerning other better studied cognitive processes—how they might interact and contribute to moral judgment.

In some sense, all models of moral judgment depend on some sort of "faculty" that functions to integrate the outputs of others cognitive processes (e.g., theory of mind) in order to compute a distinctly moral judgment. Then, any debate about the future of moral psychology might simply concern where to direct our empirical efforts—the moral faculty or the processes that feed into the faculty. This might turn on the complexity of the computations performed by the moral faculty: How much or how little work is done before the moral faculty runs its moral computations? Current models posit relatively simple moral rules, for example, "ME HURT YOU" (Greene et al., 2004) and "it is wrong to intentionally cause harm" (Cushman, Young, & Hauser, 2011; Mikhail, 2007). The simplicity of these moral computations recommends the worthy challenge of characterizing the messier processes that provide the inputs to the moral faculty.

It may be also worth noting, however, that while the moral rules themselves may be relatively simple, "hurt" and "harm" may require some deconstruction into further component parts (S. Carey, personal communication). What constitutes "hurt" and "harm"? Causing distress to another person (Leslie, Knobe, & Cohen, 2006)? Violating moral norms that extend beyond physical harms, to norms concerning fairness, community, authority, and purity (Haidt, 2007)? Hindering others—something that even 6-month-old infants recognize as "bad" (Hamlin, Wynn, & Bloom, 2007; Kuhlmeier, Wynn, & Bloom, 2003)? How "hurt" and "harm" are filled out may turn out to be learned or innate (or some combination), learned explicitly or associatively (Blair, 1995), culturally bound or universal. In the meantime, any innate content of "hurt" and "harm" may be one candidate for what is uniquely moral.

Another candidate, though, is the moral faculty itself, in other words, not merely the content of "hurt" and "harm" within the moral computation, but that which performs the computation. Indeed, some mechanism must take nonmoral inputs and deliver a uniquely "moral" judgment. And, yet, it is also possible that whatever integrative mechanism that takes, for example, mental states and outcomes as inputs to compute moral permissibility is no different from that which takes height and radius to compute volume (Anderson & Cuneo, 1978). What would be needed then is positive evidence for a specifically "moral" faculty. This evidence might take the form of a specific neural process dedicated to integrating information for moral judgment (Hsu et al., 2008), or the systematic transformation of the nonmoral inputs, post moral computation. For example, are there unique behavioral or neural signatures of theory of mind for moral judgment, compared to theory of mind deployed in nonmoral contexts, for predicting and explaining behavior (Knobe, 2005; F. Cushman, personal communication)? If so, then searching for a moral faculty may indeed be worth the effort.

Conclusion

In the past decade, cognitive psychology and neuroscience has started to reveal moral thinking in the mind and brain. Moral judgment includes a complex set of cognitive processes, including reason, emotion, and theory of mind, each with distinct behavioral and neural signatures. Conscious

reasoning from explicit principles or theories may lead directly to some moral judgments and allow us to correct others. Emotional processes, including disgust and empathy, may drive us to judge some actions as harmful, unfair, or impure, and also motivate us to behave prosocially. Theory of mind enables us to evaluate moral agents based not only on their actions and effects on the external world but on the internal contents of agents' minds—their beliefs and intentions. Finally, a moral faculty may serve to integrate the information from these cognitive processes and generate a uniquely moral judgment of right or wrong. The science of morality therefore requires at once investigating the multitude of cognitive and neural processes known to function in other domains, as well as exploring the possibility of processes dedicated specifically to morality.

Future Directions

1. How do different cognitive processes for moral judgment interact?
2. What is the relationship between moral judgment and moral behavior?
3. What are the differences and similarities between moral judgment of self versus other?
4. What are the differences and similarities between moral judgment of ingroup versus outgroup members?
5. Do distinct moral domains (e.g., harm versus purity) follow distinct cognitive rules?

References

Anderson, N., & Cuneo, D. (1978). The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General*, *107*(4), 335–378.

Anderson, S., Barrash, J., Bechara, A., & Tranel, D. (2006). Impairments of emotion and real-world complex behavior following childhood—or adult-onset damage to ventromedial prefrontal cortex. *Journal of the International Neuropsychology Society*, *12*, 224–235.

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, *2*, 1032–1037.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder*. Washington, DC: Amerian Psychological Association.

Barrash, J., Tranel, D., & Anderson, S. (2000). Acquired personality disturbances associated with bilateral damage to the ventromedial prefrontal region. *Developmental Neuropsychology*, 18, 355–381.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52(2), 336–372.

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

Blair, R. J. R. (1997). Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences*, *22*, 731–739.

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*(2), 84–92.

Cima, M., Tonnaer, F., & Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Social Cognitive and Affective Neuroscience*, *5*(1), 59–67.

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

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PLoS One*, *4*(8), e6699.

Cushman, F., Young, L., & Hauser, M. (2011). Patterns of moral judgment derive from non-moral psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Cushman, F., 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.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45.

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*(3), 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*, 998–1002.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personal and Social Psychology*, *65*(4), 613–628.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hare, R. D. (1991). *The hare psychopathy checklist-revised*. Toronto, ON: Multi-Health Systems.

Harenski, C. L., & Hamaan, S. (2006). Neural correlates of regulating negative emotions related to moral violations. *Neuroimage*, *30*(1), 313–324.

Hart, H. L. A. (1968). *Punishment and responsibility*. Oxford, England: Oxford University Press.

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*, 1–21.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H. P., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport*, *14*, 1215–1219.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095.

Inbar, Y., Pizarro, D., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, *23*, 714–725.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435–439.

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20(2), 404–410*.

Kamm, F. M. (2001). *Morality, mortality: Rights, duties, and status*. New York: Oxford University Press.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*, 357–359.

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

Kohlberg, L. (1981). *Essays on moral development, Vol. 1. The philosophy of moral development*. New York: Harper Row.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402–408.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: "Theory of mind" and moral judgment. *Psychological Science*, *6,* 421–427.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*, 273–286.

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

Mendez, M., Chen, A., Shapira, J., & Miller, B. (2005). Acquired sociopathy and frontotemporal dementia. *Dementia and Geriatric Cognitive Disorders*, *20*(2–3), 99–104.

Mikhail, J. M. (2002). *Aspects of a theory of moral cognition. Unpublished Public Law and Legal Theory Research Paper Series*,Georgetown University Law Center, Washington, DC.

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

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*(4), 655–663.

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *Neuroimage*, *16*, 696–703.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiulo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, *22*, 2730–2736.

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignacio, F. A., Bramati, I. E., Caparelli-Daquer, E. M., & Eslinger, P. J. (2005). The moral affiliations of disgust. *Journal of Cognitive Behavioral Neurology*, *18*(1), 68–78.

Moll, J., Eslinger, P. J., & de Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task: Preliminary functional MRI results in normal subjects. *Arq Neuropsiquiatr*, *59*(3-B), 657–664.

Moran, J., Young, L., Saxe, R., Lee, S., O'Young, D., Mavros, P., & Gabrieli, J. (2011). Impaired theory of mind for moral judgment in high functioning autism. *Proceedings of the National Academy of Sciences*, *108*, 2688–2692.

Ongur, D., & Price, J. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys, and humans. *Cerebral Cortex*, *10*, 206–219.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, *1*(3–4), 245–258.

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

Rolls, E. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, *3*, 284–294.

Saver, J. L., & Damasio, A. (1991). Preserved access and processing of social knowlede in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, *29*, 1241–1249.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *Neuroimage*, *19*(4), 1835–1842.

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, *80*(4), 1197–1209.

Schaich Borg, J., Lieberman, D., & Kiehl, K. A. (2008). Infection, incest, and iniquity: Investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, *20*(9), 1529–1546.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personal and Social Psychology Bulletin*, *34*(8), 1096–1109.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia*, *45*(13), 3054–3067.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476.

Vollm, B. A., Taylor, A. N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., … Elliott, R . (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage*, *29*(1), 90–98.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Children Development*, *72*(3), 655–684.

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

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*, *65*(6), 845–851.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences USA*, *107*(15), 6753–6758.

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

Young, L., & Koenigs, M. (2007). Investigating emotion in moral cognition: A review of evidence from functional neuroimaging and neuropsychology. *British Medical Bulletin*, *84*, 69–79.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology, 1(3), 333–349*.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*, 1912–1920.

Young, L., & Saxe, R. (2009a). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405.

Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*(5792), 1451–1452.