CHAPTER 19 Moral Judgment

Michael R. Waldmann, Jonas Nagel, *and* Alex Wiegmann

Abstract

The past decade has seen a renewed interest in moral psychology. A unique feature of the present endeavor is its unprecedented interdisciplinarity. For the first time, cognitive, social, and developmental psychologists, neuroscientists, experimental philosophers, evolutionary biologists, and anthropologists collaborate to study the same or overlapping phenomena. This review focuses on moral judgments and is written from the perspective of cognitive psychologists interested in theories of the cognitive and affective processes underlying judgments in moral domains. The review will first present and discuss a variety of different theoretical and empirical approaches, including both behavioral and neuroscientific studies. We will then show how these theories can be applied to a selected number of specific research topics that have attracted particular interest in recent years, including the distinction between moral and conventional rules, moral dilemmas, the role of intention, and sacred/protected values. One overarching question we will address throughout the chapter is whether moral cognitions are distinct and special, or whether they can be subsumed under more domain-general mechanisms.

Key Words: moral psychology, moral judgment, norms, moral domains, intention, folk psychology, emotion, reasoning, cross-cultural psychology, neuroscience, trolley problem, convention, protected and sacred values, heuristics and biases, dual-process theory, moral grammar, side-effect effect

Introduction

The past decade has seen a renewed interest in moral psychology. Empirical research on morality is not new, of course. There has been a long tradition in different fields, such as social and developmental psychology. Nevertheless, a unique feature of the present endeavor is its unprecedented interdisciplinarity. For the first time, cognitive, social, and developmental psychologists, neuroscientists, experimental philosophers, evolutionary biologists, and anthropologists collaborate to study the same or overlapping phenomena.

In this review, we will focus on research trying to elucidate the cognitive and affective foundations of *moral judgment*. As a first approximation one can say that moral judgments refer to the rightness or wrongness of specific acts or policies. A central question that will be repeatedly brought up in this review is whether we need a separate field of moral psychology to study this specific class of judgments. Do moral judgments possess characteristics that make them qualitatively distinct from other judgments? This question can be divided into two subquestions: *(1)* Are there moral rules people universally invoke when making moral judgments, and *(2)* Are moral cognitions a natural kind with specialized cognitive machinery, or are moral judgments just a special case of judgments in general? We will try to answer these two questions by reviewing recent studies investigating whether moral rules are universal and whether there is evidence for an innate module devoted to moral cognitions.

The answer to the first question seems to be no. Cross-cultural research has made it clear that although a variety of moral rules, such as "do no harm," are strongly endorsed in Western cultures, at least by liberals, they are not universally endorsed (see Rai & Fiske, 2011). For example, whereas in some societies hitting and fighting is impermissible, in other cultures certain forms of violence are praised. Some cultures allow violence only to outgroup members; others also encourage violence within their ingroup and find it acceptable that children, women, or animals are harmed in some circumstances. Other moral domains, such as concerns about sexuality, fairness, health, or food, are also highly variable (see Prinz, 2007; Rai & Fiske, 2011; Sripada & Stich, 2006). Thus, the contents of moral rules vary widely across cultures.

One possibility to address the second question is to look at evidence showing that moral cognitions are innate. The evidence about lack of universality already indicates that it is unlikely that specific moral rules (e.g., "do no harm") are innate. Although some researchers argue that some moral rules may be components of a universal moral grammar (Hauser, 2006; Mikhail, 2011), the evidence for this claim is weak (see Prinz, 2007; also see section on "Moral Grammar Theory"). Some researchers have therefore proposed that moral cognitions are nothing special: Moral reasoning is just domain-general reasoning with moral contents. Bucciarelli, Khemlani, and Johnson-Laird (2008) have, for example, claimed that moral reasoning can be modeled as deontic reasoning with the contents of the rules determined by cultural norms. Similarly, Gigerenzer (2010) claims that there is no special class of moral heuristics. Instead the same domain-general social heuristics guide moral and nonmoral behavior (e.g., "If there is a default, do nothing about it"). Although there can be no doubt that domain-general processes influence moral reasoning (see section on "Domain-General Cognitive Theories"), it also seems implausible to fully reduce moral rules to deontic rules. Deontic rules do not differentiate between moral rules and mere, arbitrary conventions, which seem psychologically distinct (see section on "The Moral/Conventional Distinction").

Currently there is a debate about whether the evidence favors the theory that we are innately disposed to acquire *moral rules*, which share a core content that may be variably instantiated in different cultures (Hauser, 2006; Joyce, 2006), or whether we are simply disposed to acquire *norms* whatever their content may be. Sripada and Stich (2006) differentiate norms from both moral rules and mere conventions. Like moral rules, norms typically transcend mere conventions and are considered independent of what an authority says. People believe that the norms they follow should be honored as ends, not as means to achieve a goal. Moreover, norm violations often lead to punitive emotions, such as anger or guilt. However, people who endorse norms do not necessarily claim universality for them, a feature typically associated with moral rules (see section on "The Moral/Conventional Distinction"). Sripada and Stich (2006) have suggested a domain-general norm acquisition mechanism that is capable of acquiring norms, including moral ones, the specific contents of which are culture-dependent.

This dispute is hard to settle because it depends on how moral norms are distinguished from norms in general. The review by Machery and Mallon (2010) concludes that currently the most parsimonious account is that people are universally disposed to acquire norms in general. Moral norms are then a special case; their contents are specified by the culture in which a person is born.

The question whether there is a cognitive module devoted to moral cognitions is also complicated by the fact that moral concerns may be subdivided into different domains. Extending Shweder, Much, Mahapatra, and Park's (1997) theory, Haidt and Joseph (2007) propose five moral domains that are characterized by unique adaptive challenges, contents, triggering stimuli, virtues, and emotions. In Western cultures concerns with Harm/Care and with Fairness/Reciprocity dominate. Harm/Care concerns are triggered by suffering and distress, especially by one's kin, and are accompanied by the emotion of compassion. The Fairness/Reciprocity domain deals with cheating, cooperation, and deception and is accompanied by the emotions anger, gratitude, and guilt. However, there are further domains. Ingroup/Loyality norms regulate group cooperation through pride and anger, whereas Authority/Respect norms control hierarchies by recruiting the emotions respect and fear. Finally, many cultures are concerned with Purity/Sanctity, which consists of norms referring to food, health, and sexuality (thus conceiving the body as sacred), often enforced through feelings of disgust. These moral values are not only needed to understand other cultures, but there are also differences within the Western culture. For example, conservatives are more likely than liberals to embrace all these values. In contrast, Western liberals mainly emphasize Harm and Fairness/Justice-based concerns (Graham, Haidt, & Nosek, 2009; see also Wright & Baril, 2011).

Whereas Haidt and Joseph (2007) believe that these five domains correspond to adaptations that led to innately specified dispositions to acquire domain-specific moral norms, Rai and Fiske (2011) present an alternative theory that views norms as mechanisms to regulate specific types of social relations. Unity is the motive to care for and support the integrity of ingroups, Hierarchy is the motive to respect rank in social groups, Equality is the motive for balanced in-kind reciprocity, and Proportionality is the motive for rewards and punishments to be proportionate to merit and judgments to be based on a utilitarian calculus of costs and benefits (i.e., market pricing). People are simultaneously parts of several social relations so that moral norms may vary in different contexts. For example, harm may be prohibited within ingroups that rely on the implicit assumption of Equality, whereas it may be obligatory when it is proposed in the context of Hierarchy (e.g., war) or Proportionality (e.g., punishment). This theory is consistent with the assumption that some moral norms are innate, but innateness claims here are not made about the norms but rather about the universality of specific types of social relations, where moral cognition is still part of broader social-relational cognition.

Our review of research on moral judgment will start with a critical discussion of competing theories of moral judgment. We will present these theories along with selected experimental behavioral and neuroscientific studies supporting the respective theory. We will then discuss these theories in the context of a selected number of specific research topics that have attracted particular interest in recent years.

Critical Review of Global Theories

In this section we will review and discuss global theories of moral judgment. Although these theories typically are presented as general frameworks, it will turn out that the focus of the theories differs.

Kohlberg's Rationalist Theory

Kohlberg's (1981) important theory of moral development, which was inspired by Piaget's (1932) view is discussed in many current accounts as an example of a theory that views conscious moral reasoning as a central component of morality (Haidt, 2001; Hauser, 2006). Kohlberg's (1981) famous method to study moral competencies was simple. He presented subjects (mainly children and adolescents) with dilemmas in which different moral factors conflicted. For example, in the famous *Heinz dilemma*, Heinz's dying wife can only be saved by taking a new drug that a pharmacist has developed. The production of the drug costs $200, but the pharmacist charges $2,000, double of what Heinz can pay. The pharmacist refuses to sell the drug cheaper so that Heinz eventually decides to break into the pharmacy and steal the drug. Kohlberg asked his subjects whether Heinz should have done this. He was primarily interested in the justifications for the answers, which he coded to reconstruct the level of moral reasoning.

Kohlberg found that children from many cultures typically move through a sequence of levels and substages, although not everyone reaches the higher levels of reasoning (see also Crain, 1985). Level 1 represents *preconventional* morality. This level is characterized by an orientation toward the likely punishment or obedience toward fixed rules ("do not steal"). In Level 2, the level of *conventional morality*, typically reached in adolescence, values of family and society come into play. Here the children think that people should live up to the expectations of family and society, and be good persons. For example, Heinz's behavior might be defended as good, whereas the pharmacist might be described as greedy. Later within Level 2, subjects become more concerned with society as a whole with an emphasis on laws, respecting authorities, and performing duties to maintain social order. In Level 3, *postconventional* morality, the justifications transcend the level of norms and laws and focus on the legitimacy of the norms regulating society. In this stage violations of individual rights, such as liberty and life, may be invoked to justify behavior that breaks the law.

Kohlberg did not believe in innate factors driving moral development but rather viewed the transition between levels as driven by the opportunities afforded in everyday social interactions. Change may occur as a result of everyday role taking and perspective change fostering empathy, or it may be driven by reflections about moral situations.

DISCUSSION

Kohlberg was a rationalist. He believed that our moral judgments are driven by reasoning processes, and that progress in moral development is driven by reflections and discussions. Many current theories criticize this rationalist assumption. For example, Haidt (2001) argues that moral intuitions are primarily based on unconscious intuitions, with justifications being post hoc rationalizations (see section on "Haidt's Social Intuitionist Model"). Thus, it can be questioned whether the justifications Kohlberg elicited really caused the intuitions in the moral dilemmas. Other researchers acknowledge that occasionally moral intuitions may be based on reasoning processes, but they argue there are also important cases in which we do not have conscious access to the factors driving our intuitions (see Cushman, Young, & Hauser, 2006). A related critique is that Kohlberg's focus on levels and stages underestimates the context dependency and variability of moral reasoning. Moral intuitions in different cases may be driven by different context factors so that the reduction to a general level may be an oversimplification.

Kohlberg's theory has also been criticized as culturally biased (see also Gilligan, 1982, for the claim that Kohlberg's higher levels are biased toward the reasoning of males). Kohlberg argues that people in all cultures go through the same levels, but there may be differences in the rates of development and the end state. For example, he found that in urban contexts people typically reach Level 2 and some lower stages of Level 3, whereas in tribal communities and small villages, Level 1 is rarely surpassed. Simpson (1974) argues that Kohlberg has developed a stage model based on the Western philosophical tradition and has then imposed it on non-Western cultures. Kohlberg's response was that his theory is not about the specific values different cultures endorse but about general modes of reasoning, but this position has become increasingly questionable in light of the diminished role of justifications as evidence for morality in current theories.

Emotion-based Theories

Kohlberg's (1981) theory can be traced back to rationalist philosophy. Moral reasoning is described as conscious deliberation; the sequence of moral stages seems to lead toward ethical positions that have been elaborated in Kant's (1959) and Rawls' (1971) philosophy. Many current theories are instead influenced by Hume's (1960) philosophy of morality. Hume held the view that moral distinctions are not results of reasoning processes but can be derived, analogous to aesthetic judgments, from affectively laden moral sentiments: feelings of approval and disapproval felt by spectators who contemplate a character's trait or action.

HAIDT'S SOCIAL INTUITIONIST MODEL

Inspired by Hume, Haidt (2001) defines moral judgments as "evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture" (p. 817). In his model an important distinction is between reasoning, a conscious activity in which conclusions are derived through several steps, and intuition, also a cognitive process that is characterized "as the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion" (p. 818). Thus, whereas reasoning is largely conscious, intuition is based on automatic, unconscious processes.

In his social intuitionist model, the primary link underwriting moral judgments is the link between the eliciting situations and moral intuitions. Reasoning processes may modify judgments, but in the model they are optional and start after initial intuitions have been formed. The role of reasoning is often to provide post hoc rationalizations of the already formed moral intuitions. Occasionally, private reasoning may override the initial intuitions, but this is relatively rare. Apart from these processes that occur within the individual, the model also contains links to other members of the social group. Other people may be influenced by moral judgments, or they may change their minds on the basis of discussions. Moreover, the intuitions and reasoning of others may influence the moral intuitions of the individual. Thus, individuals are embedded in social contexts and their norms.

Evidence for the existence of a direct intuitive link comes from a number of studies (see Haidt, 2001, 2007; Haidt & Kesebir, 2010). In several studies about harmless taboo violations (e.g., eating a pet dog; a consensual incestuous relation with birth control), many subjects judged that the acts were morally wrong but were unable to provide reasons for their judgments (i.e., moral dumbfounding; see Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993). Moral dumbfounding can be viewed as evidence for the unconscious elicitation of moral intuitions.

The automaticity of moral judgments may lead to misattributions of consciously accessible affects. For example, Wheatley and Haidt (2005) selected highly hypnotizable subjects who were given a posthypnotic suggestion to feel a flash of disgust when they read an arbitrary word. Moral vignettes were presented that did or did not contain the word. The results showed that moral judgments can be made more severe by the presence of the hypnotically triggered disgust. In a related study, Schnall, Haidt, Clore, and Jordan (2008) manipulated the context in which subjects made moral judgments about a character in a story. Subjects who scored high on a *private body consciousness* scale made harsher judgments in the presence of a bad smell ("fart spray") than in its absence. Similarly, Eskine, Kacinik, and Prinz (2011) showed that the taste of a beverage influences moral judgments about other people.

Although our focus in the present chapter is on theories of moral reasoning, it should be noted that Haidt is a social psychologist who is mainly interested in processes that go beyond individual reasoning (see Haidt, 2007; Haidt & Kesebir, 2010). Individuals are embedded in large social contexts in which they influence others, as they are influenced by others. Thus, whereas many researchers have used the model of a lonely "intuitive scientist" to study moral reasoning, Haidt prefers the metaphor of an "intuitive politician." The focus on larger social contexts also highlights the important role of group norms, cooperation, and methods of the society to punish defectors, which are often neglected in research centered on judgment and decision making (see also Haidt's theory of moral domains in the "Introduction" to this chapter).

Discussion

Although Haidt's primary interest is social and culture psychology, we will focus here on his ideas about moral judgment. His approach proved a valuable contrast to the rationalist approaches of Kohlberg (1981). Numerous findings show that moral judgments can be intuitive and automatic. However, from the viewpoint of cognitive psychology, the contrast between reasoning and intuition excludes important possibilities in the middle. In modern theories of reasoning, it is rarely assumed that the steps undertaken by the reasoner are fully accessible to consciousness (see Harman, Mason, & Sinnott-Armstrong, 2010). Mental model theory, Bayesian theories, and even mental logic theory postulate various processes that work below the threshold of conscious awareness.

A further problem is that the step between eliciting situation and intuition remains largely opaque in Haidt's (2001) theory. Although Haidt acknowledges that intuitions are based on cognitive processes, and Haidt and Kesebir (2010) even mention that heuristics may play an important role, there is no worked-out theory of how specific situations lead to particular moral intuitions. A cognitive-affective theory of moral intuitions needs to specify how moral scenarios are perceived and categorized, and how these initial appraisals are further processed. Of course, the information processing steps and representations leading to moral intuitions may well be unconscious or only partly conscious.

Finally, the claim that moral intuition is primary and that reasoning is secondary has led to critiques. Although it seems plausible that this relation often holds given that intuitions are based on faster processes than reasoning, there are certainly also cases in which people do not have clear initial intuitions and arrive at their judgments after careful deliberation (see Haidt & Kesebir, 2010; Paxton & Greene, 2010).

THE PLACE OF EMOTIONS IN MORAL JUDGMENTS

In Haidt's (2001) theory, affective evaluations play an important role in moral intuitions, but the exact role of them is left open. In fact, virtually every theory of moral reasoning acknowledges that emotions are an important part of our moral judgments. Even Kant (1959), in his rationalist philosophy of morals, claims that moral judgments are typically accompanied by moral feelings. What is debated is the exact place of emotions in moral judgments.

Different positions can be distinguished (see Hauser, 2006; Huebner, Dwyer, & Hauser, 2009; Prinz & Nichols, 2010). The Kantian approach, which postulates that deliberate conscious reasoning processes generally precede emotions, is refuted by the findings discussed in the section on "Haidt's Social Intuitionist Model." Numerous studies have shown that moral judgments are often immediately triggered without extensive reflections.

A second possibility, in the tradition of Hume's ideas, views moral judgments as caused by distinct prior emotions. The problem with this approach is that it is unclear which emotions trigger distinctly moral judgments and how these emotions are caused. For example, feelings of disgust may alter moral evaluations (Schnall et al., 2008; Wheatley & Haidt, 2005), but not all feelings of disgust lead to moral judgments. Also there is no clear unambiguous relation between affects and judgments. Feelings of pity and compassion may occur when we observe immoral torture but also when we watch a lifesaving amputation of a leg. Thus, it seems more likely that affects, such as disgust, moderate moral evaluations that are already independently triggered by signaling the degree of aversiveness or disutility.

These problems have led Nichols (2004) to his "sentimental rules" theory. Nichols argues that moral judgments are fed by two components: a normative theory and a system of emotions. The normative theory specifies the content of moral rules that are acquired in a specific sociocultural environment; the emotions alter their character, which includes being considered serious and authority independent (see section on "The Moral/Conventional Distinction"). Due to the presence of emotions, moral rules acquire their force and impact (i.e., emotion-backed rules). Moral judgments in the absence of emotions are possible but rare in healthy subjects. Such judgments would also not have the same force and strength as judgments based on emotion-backed rules. Prinz (2007) proposes a related theory but questions the possibility of separating emotions from moral judgments. In his view, moral concepts, such as "moral" or "immoral," contain emotions as essential components (as in Hume's account).

The present research does not allow us to empirically decide between these positions (see Huebner et al., 2009). Studies in which affect was manipulated prior to or simultaneous with the scenarios (e.g., Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005) show that affect influences judgments; they do not demonstrate that emotions are necessary for moral judgments. They might instead affect the interpretation of the scenario, the evaluation of the outcomes, or the interpretation of the test question.

The results of neuroimaging and neuropsychological studies are also ambiguous. Neuroimaging studies have shown that emotional responses are integral components of moral reasoning. For example, an increased activity in the frontal polar cortex (FPC) and medial frontal gyrus was seen in moral judgments compared to judgments of nonmoral claims (Moll et al., 2002, Moll, de Oliveira-Souza, & Eslinger, 2003; see also Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).Similarly, neuropsychological studies provide strong evidence for the role of emotions in morality. For example, frontotemporal dementia patients, who suffer from the deterioration of their prefrontal and anterior temporal cortex, show blunted emotions, disregard for others, and a willingness to engage in moral transgressions (Damasio, 1994; Mendez, Anderson, & Shapira, 2005). However, the exact functional role of emotion remains unclear.

The most interesting evidence comes from a study by Koenigs et al. (2007; see also Ciaramelli, Muccioli, Làdavas, & di Pellegrino, 2007). Damage to the ventromedial prefrontal cortex (VMPC) leads to an emotional flattening and to a decreased ability to anticipate rewards and punishments (Damasio, 1994). Koenigs and colleagues showed that VMPC patients were generally able to evaluate moral dilemmas in which a victim needs to be sacrificed to save others like healthy subjects, but they differed in high-conflict dilemmas, which cause strong emotional responses in healthy subjects. Whereas these healthy subjects were primarily led by their affective responses, the VMPC patients opted for a consequentialist resolution, which simply compared numbers of victims (see also sections on "Dual-Process Theory," "Moral Grammar Theory," and "Moral Dilemmas"). Recently, Bartels and Pizarro (2011) extended these findings by showing that healthy participants who had higher scores on measures of Machiavellianism, psychopathy, and life meaninglessness indicated greater endorsement of utilitarian solutions. Again these studies show that in healthy subjects emotions influence judgments, but it is unclear how. Huebner and colleagues (2009) suggest that emotions in these cases may influence the interpretation of the scenario. Emotionally salient outcomes may be downplayed by the VMPC patients, and this in turn affects the moral evaluations.

Another strategy to elucidate the role of emotions in moral judgments is to take a closer look at the emotions that accompany judgments (see also Prinz & Nichols, 2010). The philosopher Williams (1985) has distinguished "thin" (e.g., good, bad) from "thick" moral concepts that are loaded with content (e.g., cruelty, courage). Similarly, in emotion research one can study thin affects (e.g., positive, negative) that may occur in the absence of awareness of the triggering conditions, or emotions can be described as thick relational concepts that have moral content and trigger specific moral behavior. This position views emotions as cognitively entrenched (Lazarus, 1991), and it is consistent with the views that moral emotions are constitutive of moral judgments (Prinz, 2007) or are strongly attached to moral rules (Nichols, 2004). Examples of moral emotions are anger and guilt. According to Prinz and Nichols (2010), anger is typically elicited by a violation of somebody's autonomy and often motivates retaliatory acts. In contrast, guilt is elicited by the feeling of direct or indirect responsibility for somebody's harm, especially when the harmed person belongs to the ingroup so that there is a threat of separation or exclusion. Both emotions not only express affects but are constitutive for the expression of specific moral values.

Dual-Process Theory

All theories we have discussed so far acknowledge that both conscious reasoning and emotion-based intuitions play an important role in moral judgments. They differ, however, in what process they consider primary. A further theory, the dual-process model proposed by Greene and colleagues, claims that our brains contain multiple systems driving moral intuitions, one devoted to rational, the other to emotional processes. The system underlying rational deliberations is slow, effortful, and controlled, whereas the affective system consists of automatic, largely unconscious, intuition-based processes (see Cushman, Young, & Greene, 2010, for an overview).

Initial evidence for the dual-process theory comes from a neuroimaging study by Greene et al. (2001). This study investigated moral dilemmas, such as the trolley dilemma (Foot, 1967; Thomson, 1985), which will be more extensively discussed in the section on "Moral Dilemmas." The basic scenario in trolley dilemmas describes a runaway trolley threatening to kill five people on the track. Many people find it permissible to flip a switch that redirects the train onto a side track where only one person would be killed (bystander version), whereas it is generally considered impermissible to shove a person onto the tracks to stop the train (footbridge version), despite the fact that in both scenarios one person would be killed instead of five (Hauser, Cushman, Young, Jin, & Mikhail, 2007; see also Fig. 19.1). Greene and colleagues argued that the footbridge case is more *personal* and therefore triggers a negative affective response, consistent with deontological philosophy, whereas the bystander version is more *impersonal*, which therefore leads to a consequentialist weighing of lives harmed. Thus, the theory is that there are two separate systems in the brain, one triggering affect-based, and the other rational, consequentialist responses. The nature of the moral dilemma decides which system predominates. Supporting this theory, the results of Greene at al. (2001) indeed showed that brain regions associated with emotions, such as the medial prefrontal cortex, were more active with personal dilemmas, whereas brain regions associated with controlled cognitive processes such as working memory and abstract reasoning were more active with impersonal dilemmas. Greene et al. (2001) also reported reaction time data favoring their theory, but a reanalysis by McGuire, Langdon, Coltheart, and Mackenzie (2009) shows that outliers that needed to be removed are mainly responsible for the observed pattern.

In follow-up studies, Greene and colleagues presented particularly difficult "high-conflict" dilemmas (see also "The Place of Emotions in Moral Judgments"). For example, in one example a situation is presented in which the father of a crying baby can only save his other children from enemy soldiers if he smothers his crying child to death; otherwise all children including the crying one would be killed. Greene, Nystrom, Engell, Darley, and Cohen (2004) have shown that in such cases the anterior cingulate cortex (ACC) is active, which is involved when incompatible responses are activated. Moreover, consistent with the dual-process theory, the dorsolateral prefrontal cortex was more active when what Greene et al. viewed as the "consequentialist" response (i.e., killing the baby) was given. In a related study, Greene, Morelli, Lowenberg, Nystrom, and Cohen (2008) presented subjects with such hard moral dilemmas while at the same time exposing them to a cognitively demanding secondary task. The results showed that only the "consequentialist" responses were slowed down by this procedure, which had no effect on the affect-based (according to Greene et al. "deontological") responses. Similarly Suter and Hertwig (2011) manipulated how much time they granted their subjects to render a judgment in order to constrain controlled processes. They found more "deontological" responses under strict time constraints than in the contrasting condition in which subjects had more time to respond.

Whereas these findings showed selective interference with the assumed consequentialist brain area, the studies with patients with VMPC lesions discussed in "The Place of Emotions in Moral Judgments" are evidence for interference with the emotional areas. These patients tended to give the "consequentialist" answer in high-conflict dilemmas, while the emotion-based response was blunted (Koenigs et al., 2007; Mendez et al., 2005).

DISCUSSION

Multisystem theories are attractive in many areas because they integrate a large body of different findings compared to single-system theories, which are less flexible. However, the particular version of a dual-system theory by Greene and colleagues has drawn a number of critiques. Prinz (2008) argues that the results of Greene et al.'s (2001) study are consistent with the view that the dilemmas trigger a strong negative emotional response to harming an innocent bystander and a weaker positive emotional response to saving five potential victims. Differences in salience of the harming versus the saving options may explain the observed patterns. Moll and De Oliveira-Souza (2007) suggest that differential activations of brain areas underlie different prosocial emotions, which integrate emotional and cognitive processes rather than putting them in conflict.

A further possible critique concerns the interpretation of the contents of the two postulated brain areas. Although the exact characterization shifts across publications (see Cushman et al., 2010), the brain areas identified as underlying moral judgments were often characterized using labels describing philosophical positions, such as deontological and consequentialist (see Greene, 2008). However, it is questionable whether different brain areas embody complex contentful moral philosophies rather than more domain-general processes, such as affective reactions versus rational deliberations. Moreover, Kahane et al. (in press) have pointed out a confounding: In the experimental scenarios used by Greene and colleagues (2001, 2004) the deontological option is always also the more intuitive one. Thus, the discovered asymmetries between responses corresponding to deontological vs. consequentialist rules might in fact be due to differences in intuitiveness. Using additional scenarios in which the consequentialist response is more intuitive than the deontological one (e.g., lying in order to prevent serious harm), Kahane and colleagues demonstrated that characteristic differences in neural activation are more closely related to the intuitiveness of the response options than to their deontological versus consequentialist content.

Furthermore, the characterization of responses based on numbers of saved lives as consequentialist overstates the finding. For example, the people who judge killing the baby to be acceptable in the crying baby dilemma outlined earlier need not have applied a consequentialist moral philosophy; the same conclusion could have been reached by application of a deontological rule (Kamm, 2007) or without reference to any formal moral theory (see also Kahane & Shackel, 2008). So far, there is no evidence that a version of an intuitive consequentialist theory is coded anywhere in the brains of naïve subjects. It would be more parsimonious to say that in different conditions different aspects of the scenarios (for example, acts versus number of victims) are highlighted (cf. Bartels, 2008). This interpretation would also have the advantage that VMPC patients and psychopaths would not have to be viewed as particularly rationalist, consequentialist reasoners (see also Bartels & Pizarro, 2011).

Finally, there are different versions of dual-process theories, which seem equally consistent with the data (see Evans, 2007; Evans, Chapter 8). One possibility is the theory endorsed by Greene and colleagues, which assumes there are two dissociable systems that operate independently, acquire different knowledge, and compete in the control of behavior. However, another possibility is that there is only a single database (e.g., trolley dilemmas with various features, such as acts and outcomes) and sequential processes operating on these representations. The initial fast processes may lead to heuristic or emotion-based judgments, whereas in some circumstances the output of the initial pass is further processed by more effortful, controlled processes (Evans, 2007; Kahneman & Frederick, 2005). This theory would also explain the findings without the need to postulate multiple brain areas embodying different moral philosophies.

Moral Grammar Theory

The theories we have discussed so far largely focus on whether moral reasoning is driven by intuitive affective processes or by conscious reasoning. None of the theories specifies a precise computational mechanism that translates situational input into moral judgments. It is the contribution of Mikhail (2007, 2011 to rise to the challenge and present a sketch of such a computational theory (see also Hauser, 2006). Like the other modern theorists, Mikhail accepts that moral judgments are typically not based on conscious deliberate reasoning. However, this does not mean that the underlying processes cannot be reconstructed as steps of computational information processing. Most cognitive theories, in both higher and lower order cognition, assume the operation of unconscious processes underlying the mental products that rise to consciousness.

Mikhail's theory of universal moral grammar is inspired by Chomsky's (1957) linguistic grammar theory. We have intuitions about the grammaticality of sentences, which can be explained as the output of the operation of a complex unconscious system of syntactic rules. Similarly, Mikhail argues that our judgments of moral permissibility may be driven by an unconscious moral grammar that contains moral rules. The moral grammar theory holds that individuals are intuitive lawyers who possess unconscious knowledge of a rich set of legal rules along with a natural readiness to compute mental representations of human acts and omissions in legally cognizable terms. Following Chomsky, Mikhail claims that the moral grammar is innate and universal. The innateness claim is defended by a variant of the *poverty of stimulus argument*, according to which the learning input of children would not sufficiently constrain the moral rules they seamlessly acquire. A further argument is that people often have clear moral intuitions without being able to verbally explicate or justify them (Cushman et al., 2006). Empirical support for the theory mainly comes from a large Internet study in which thousands of subjects from various countries were confronted with variations of the trolley and other dilemmas (see section on "Moral Dilemmas").

Moral grammar theory specifies a series of computational steps transforming the observed stimuli into morally relevant internal representations. Initially a set of conversion rules encodes the temporal structure of the presented stimulus (e.g., a trolley dilemma) and transforms it into a representation of the underlying causal structure. For example, in the bystander dilemma (see sections on "Dual-Process Theory" and "Moral Dilemmas"; see also Fig. 19.1), the temporally ordered events "throwing a switch," "turning the train," and "killing one man" are integrated into a causal chain representation. This way knowledge is acquired about morally relevant causal features, such as whether death is a side effect or a means of the proposed act or omission. Next, other conversion rules translate the causal representation into a moral representation by assigning evaluations to the effects (good vs. bad). This representation is further converted into a representation of the underlying intentional structure. In a scenario with both good and bad effects, it is by default assumed that the good outcome is the intended outcome, whereas the side effect is simply foreseen. However, if the bad effect is a means to a good outcome, it necessarily is intended, because it constitutes the only route to the good effect. Further morally relevant information is filled in, such as whether the act is a case of intentional battery or whether the victim is harmed without having given consent. Finally, a set of deontic rules is applied to the final representation of the stimulus, yielding a judgment of obligation, permissibility or prohibition of the encoded action.

So far Mikhail (2009) has focused on two deontic rules, which are particularly relevant for trolley dilemmas. The "prohibition of battery and homicide" forbids an agent to purposely cause harm to a nonconsenting victim. A second rule, the *doctrine of double effect* (DDE), can be traced back to Aquinas and to Roman Catholic theology from the 19th century. The correct interpretation of this doctrine is under dispute (see Woodward, 2001). Double effect refers to the two effects an action might have, the intended goal and a foreseen but unintended side effect. According to Mikhail's (2009) reading of the DDE, an otherwise prohibited act, such as battery or homicide, with good and bad effects may be permissible if the prohibited act itself is not directly intended, only the good outcomes are intended, the bad ones merely foreseen, the good effects outweigh the bad one, and there are no better alternatives. This rule is consistent with the finding that people typically consider it acceptable to redirect the trolley in the bystander version but oppose the act in the footbridge version. In the bystander version the act generates a bad effect as a side effect, which is not intended but only foreseen, whereas in the footbridge version a person is directly killed as a means to a greater good. Thus, this is a case of intentional battery.

DISCUSSION

Although to date the computational theory underlying moral grammar theory is only a sketch of a model, its precision and detail vastly surpass what other moral theories currently offer. The focus on processing details may also be the reason why the scope of the model is thus far limited. It is clearly developed to account for trolley dilemma intuitions, whereas it is less clear how other moral cases will be handled.

The focus on a restricted class of harm-based dilemmas and on deontic rules that are taken from Western moral philosophy (e.g., DDE) cast doubt on the claim that the theory is universally valid as claimed. Although the Internet study has collected data in numerous cultures, we will show in the section on "Moral Dilemmas" that alternative explanations of the effects are plausible. It seems questionable that a principle, such as the DDE, is universally valid. Even an initially plausible deontic principle, such as the prohibition of intentional battery, does not seem to hold universally (see "Introduction").

In some versions of the moral grammar theory, the analogy to Chomsky's (1957) grammar theory is carried even further to accommodate findings of intercultural differences. In his principles and parameter theories, Chomsky claimed that we are born with a universal innate grammar that contains fixed principles but also parameters that are set by the linguistic environment of the person. This model explains why people are able to quickly learn very different languages with differences in the syntax. Analogously, it has been argued that moral grammar may contain principles, such as the doctrine of double effect, and parameters that are set by the culture (Dwyer, 2006; Hauser, 2006; Roedder & Harman, 2010). However, no formal version of a moral grammar of this kind has been worked out yet, so this proposal remains untestable. Moreover, moral principles, such as the doctrine of double effect, combine domain-specific rules with domain-general processes (e.g., intentional and causal analyses) so that it is unclear how both types of processes are organized within an innate module (see also Cushman & Young, 2011).

There are further critiques casting doubts on the analogy between moral grammar and Chomsky's (1957) syntax theory (see also Dupoux & Jacob, 2007). First, it seems questionable to compare grammaticality judgments with permissibility judgments. Whereas with sufficient training about the meaning of the concept of syntax we know whether a sentence like "colorless green ideas sleep furiously" is grammatical, our moral intuitions, even with professional training, are extremely context sensitive and hardly ever clear cut. It seems unlikely that the factors influencing moral judgments are encapsulated in a way that warrants the modularity assumption. Moral intuitions seem to be closer to semantics and pragmatics than syntax. Moreover, the fact that people do not have conscious knowledge about moral rules does not entail innateness. There is a large literature on artificial grammar learning, for example, which similarly demonstrates judgments in the absence of valid verbal justifications (see Litman & Reber, 2005, for an overview).

These critiques do not diminish the contribution of Mikhail (2009). It is possible to work on a theory of moral rules without accepting the innateness or universality claims. In fact, the hypothesis that moral judgments are driven by moral principles, such as the doctrine of double effect, can easily be isolated from other claims (see Cushman et al., 2010) and can be tested independently (see section on "Moral Dilemmas").

Moral Heuristics

A further approach to studying moral judgments is motivated by the *heuristics and biases* paradigm, which comes in several, often competing variants (see Sinnott-Armstrong, Young, & Cushman, 2010). A general assumption underlying research on heuristics is that people use mental shortcuts or rules of thumb that generally work well but also lead to systematic errors in specific circumstances (see Sunstein, 2005). Heuristics may operate consciously or unconsciously. We will restrict our discussion here to contentful heuristics; a discussion of the role of affect, which has also sometimes been described in terms of heuristics, will not be repeated here (see section on "The Place of Emotions in Moral Judgments").

One specific characterization of the concept of heuristics describes their use as *attribute substitution* (Kahneman & Frederick, 2005). Often target attributes T of an object X are not easily accessible so that a person instead uses an attribute H, the heuristic attribute, which is correlated with X and is more accessible. The user of the heuristic tends to believe in T when H is present. This simple model applies to many cases within the heuristics and biases program. For example, it has been shown that availability (H) is used to infer probability (T) of an event (X) (Kahneman, Slovic, & Tversky, 1982). Or in a competing theory context it has been shown that recognition (H) is often used as the basis of decisions about the size (T) of cities (X) (Gigerenzer, Todd, & the ABC Research Group, 1999).

This general framework has also been applied to moral reasoning. Baron (1994, 1998) has argued that consequentialism or utilitarianism provides normatively correct answers in questions of morality. But people who are not philosophically trained do not think along the lines of these normative theories, but rather use simple heuristics that often mislead them because they lead to overgeneralizations beyond the contexts in which they provide useful advice. Baron and his colleagues have tried to show that people are not thinking according to the normative consequentialist guidelines but instead use simple heuristics.

An example of such a heuristic is the "do no harm" heuristic, which may underlie the common intuition that it is worse for a physician to kill a patient with a deadly disease then let him die by refraining from any kind of medical intervention. Consequentialist philosophers argue that these cases should be treated equivalently (Singer, 1979). Baron and colleagues have shown in a number of well-controlled experiments that people consider harmful acts worse than harmful omissions with otherwise identical, predictable outcomes (i.e., omission bias).

For example, Spranca, Minsk, and Baron (1991) found that people find it worse when somebody who wants to harm a person offers this person a food item with an allergenic ingredient than when she passively watches the person who does not know about the ingredient taking this item himself.

A related research view has been suggested by Sunstein (2005), who subscribes to a weak nonutilitarian form of consequentialism that also might count types of acts or violations of rights as relevant consequences that need to be weighed. Otherwise the general approach is similar to Baron's. Sunstein has developed a catalog of heuristics, for example, "do no harm," "people should not engage in wrongdoing for a fee," "punish and do not reward betrayals of trust," or "do not tamper with nature" (see also Baron, 1998). This list of heuristics seems to come straight from Western deontological philosophy.

DISCUSSION

Many of the previously discussed moral theories have left the aspect of cognitive appraisal of the situation unspecified. Theories of moral heuristics represent an important step in the direction of specifying the rules that may underlie moral evaluations. However, it can be questioned whether the normative foundation of the heuristics approach holds in moral reasoning. In nonmoral tasks, such as the estimation of city sizes, the target attributes can be clearly measured and compared with the output of the heuristics. In moral domains it is far less clear what the target attributes are (see Sinnott-Armstrong et al., 2010). To evaluate a heuristic, it would be necessary to use a normative theory, and in ethics, even more than in other fields, there is no agreement about the proper normative theory. For various reasons, in psychology, consequentialism has been proposed as the yardstick for ethical judgments (Baron, 1994; Greene, 2008; Sunstein, 2005), but once we delve into the philosophical literature it becomes clear that there are various versions of consequentialist and nonconsequentialist ethics that are defensible (see Kamm, 2007; Parfit, 2011; Scanlon, 1999). For example, it is far from clear whether killing should really be viewed as equivalent to letting die (Kamm, 2007). Moreover, it can be argued that although a utilitarian cost-benefit analysis may be appropriate in small worlds with limited options, in realistic scenarios relevant information about possible outcomes, probabilities, and costs and benefits is simply not available to make such a complex strategy reliably applicable (Bennis, Medin, & Bartels, 2010; Binmore, 2008; Gigerenzer, 2010).

Once we abandon the commitment to a specific normative ethical position, the distinction between heuristic and target attribute breaks down. A heuristic, such as "do no harm," may then be better framed as part of a deontological target theory that people happen to endorse (Sinnott-Armstrong et al., 2010). Instead of classifying simple moral rules as heuristics for a target attribute, it may be more productive to empirically study them as building blocks underlying moral judgments. It will probably turn out that one-sentence rules are too simple to explain the many subtle context effects that are known. For example, the research on the trolley dilemma shows that people do not generally invoke a "do no harm" rule; and even a more complex rule, such as the doctrine of double effect, does not provide a complete theory (see section on "Moral Dilemmas"). The research shows that far more complex intuitive theories underlie appraisal processes than the heuristics approach suggests. Moreover, complex intuitive systems of rules may only be one possible way to represent moral knowledge. Other possibilities include memory for exemplars (e.g., of moral transgressions) or prototype representations (see Harman et al., 2010; Sripada & Stich, 2006).

Domain-General Cognitive Theories

We have discussed several theoretical approaches that model the cognitive and emotional factors underlying moral judgments. However, another possible research strategy is to treat moral judgments simply as a special case of domain-general cognitive processes. For example, various researchers have shown that principles found in behavioral economics and psychological judgment and decision theory (JDM) also affect intuitions about moral scenarios (Rai & Holyoak, 2010; Reyna & Casillas, 2009). Examples include framing effects (Bartels & Medin, 2007; Kern & Chugh, 2009; Petrinovich & O'Neill, 1996; Sinnott-Armstrong, 2008), outcome bias (Gino, Shu, & Bazerman, 2010), or effects of joint versus separate evaluation (Bartels, 2008; Lombrozo, 2009; Paharia, Kassam, Greene, & Bazerman, 2009).

Another promising class of theories applicable to moral reasoning is causal model theory. Moral judgments are generally concerned with the evaluation of acts that lead to direct and indirect effects. It has been shown that the locus of intervention, the intentions causing the acts, and the causal structure leading from acts to good and bad effects affect judgments (Cushman & Young, 2011; Sloman, Fernbach, & Ewing, 2009; Waldmann & Dieterich, 2007).

Others have pointed out the impact of attentional processes in moral judgment. As Bartels and Medin (2007; Bartels, 2008; Sachdeva & Medin, 2008) have demonstrated, moral scenarios may be evaluated very differently depending on where subjects' attention is directed (see also section on "Sacred/Protected Values"). In a related vein, Waldmann and Wiegmann (2010) argued that aspects of the causal structure of moral dilemmas may affect moral judgment by influencing people's attentional focus on alternative counterfactual contrasts (see section on "Moral Dilemmas").

The list of domain-general factors influencing moral judgment is much longer still. Abstract, high-level construal of actions leads to amplified ascriptions of both blame and praise to agents compared to more low-level, concrete representations of the same actions (Eyal, Liberman, & Trope, 2008). Metacognitive experiences like processing fluency are also used as input for moral judgment (Laham, Alter, & Goodwin, 2009; Rai & Holyoak, 2010). Approach- and avoidance-based motivational systems seem to have similar effects on judgments about moral issues as they do in other domains (Janoff-Bulman, Sheikh, & Hepp, 2009). Extensive mental simulation, induced by closed eyes during the judgment process, makes moral judgments more extreme (Caruso & Gino, 2011). And recently, more and more researchers have pointed out the importance of individual differences in moral judgment, including need for cognition (Bartels, 2008; Bartels & Pizarro, 2011), working memory capacity (Moore, Clark, & Kane, 2008), sensitivity to reward and punishment (Moore, Stevens, & Conway, 2011), and personality traits such as extraversion (Feltz & Cokely, 2009).

Specific Research Topics

After having outlined the main theoretical approaches to the study of moral judgment, along with relevant empirical evidence taken as support of them, we now turn to four selected empirical research areas that have attracted much attention in the recent years. The aim of this section is to demonstrate the complexity of explaining moral judgments in specific tasks.

The Moral/Conventional Distinction

One of the central controversies in the field of moral psychology concerns the question of whether morality constitutes an independent domain with specific norms. Do humans reason qualitatively differently about moral rules as opposed to mere social conventions? This question has motivated various studies and led to controversial discussions.

The moral/conventional distinction was introduced to psychology by Turiel (1983). According to him, the purpose of conventional rules is to coordinate behavior in social systems. They gain their binding status by consensus within a given society, and they are arbitrary in that different agreements could have led to alternative conventions which would be just as feasible or appropriate. This implies that it is impossible to know whether a given action is in accordance with present social conventions by looking at the *action itself*. To know that it is appropriate to address your teacher by her last name, for example, you need to know that your society agreed that this is the proper way to behave since there is nothing intrinsically bad about addressing your teacher by her first name.

By contrast, the main distinguishing feature of moral rules, according to Turiel (1983), is that they are *not* arbitrary in the way conventions are. The reason for this is that they are concerned with certain *contents*: They regulate actions that have intrinsic consequences related to harm, fairness, or justice. A prototypical example of a moral transgression is that of a child pushing another child off a swing just because she wants to use it instead. According to Turiel, this act can be directly classified as harmful by any human observer familiar with pain regardless of the cultural context in which the act occurs.

Turiel (1983) argued that the moral and the conventional are separate domains of social knowledge, which are acquired and processed independently. In his view, moral rules can be empirically distinguished from conventional rules based on what Kelly, Stich, Haley, Eng, and Fessler (2007) called *signature moral pattern*. This term refers to a characteristic set of reactions people usually exhibit when they judge transgressions of prototypical moral rules: People supposedly consider moral rules, in contrast to conventional rules, to be valid even if they are not enforced by an authority (authority independence). Furthermore, moral rules are considered to be universally valid for all agents in similar circumstances across all times, places, and cultures (universality). Finally, people are expected to judge transgressions of moral rules as particularly serious and to justify their wrongness with reference to principles of harm, fairness, or justice.

An empirical paradigm to assess these signature patterns is the "moral/conventional task," in which participants are presented with someone violating a prototypical moral rule (as in the swing example) or a social convention (as in the teacher example). If Turiel (1983) is right, then cases of moral transgression should reliably elicit the signature moral response pattern, while cases of conventional transgression should elicit the *signature conventional pattern* (i.e., judgments of authority dependence, lack of universality, decreased seriousness, and justification not related to harm, fairness, or justice; see Kelly et al., 2007). This pattern of response characteristics has been confirmed in a large number of empirical studies in diverse populations, including young children and people from different cultures (see Turiel, 2006, for an overview).

Despite its doubtless merit in advancing empirical research on moral psychology, the Turiel (1983) theory faces a number of conceptual and empirical problems. On the conceptual side, Turiel's definition of morality seems to be at least in part a *petitio principii*: How people manage to recognize matters of fairness and justice seems to be just as much in need of explanation as how they recognize matters of morality. Intuitively, only the harm component seems to be a more basic concept, and more recent work trying to defend a content-based distinction mainly concentrates on harm (e.g., Royzman, Leeman, & Baron, 2009; Sousa, Holbrook, & Piazza, 2009).

The Turiel (1983) theory can also be contested on empirical grounds. In Turiel's content-based approach to defining the moral domain, harm avoidance, fairness, and justice are tacitly assumed to be universal ends. We have already pointed out in the Introduction that cross-cultural research casts doubt on this universalist assumption.

Another empirical problem has been raised by Blair (1995). He questioned the assumption that only general cognitive capacities are needed to recognize moral transgressions. He showed that incarcerated psychopaths do not respond to moral transgressions with the signature moral pattern, even though they possess all the experiential and inferential capacities that according to Turiel (1983) are sufficient to distinguish the moral from the conventional. Blair's (1995) proposal is that the human mind contains a specific module that is indispensable for this task, and which is selectively impaired in psychopathic individuals. He posits the existence of a violence inhibition mechanism (VIM) that is activated by perceptual key stimuli, mainly nonverbal facial distress cues of suffering individuals. Once activated, the VIM triggers a withdrawal response in the observer, which he or she interprets as a moral emotion. According to Blair (1995), conventional transgressions lack these VIM specific distress cues. They therefore fail to activate the VIM and lead to a different response pattern.

It is notable that on Blair's (1995) account, the main criterion for distinguishing between the moral and the conventional is shifted from a property of the *rules* (i.e., their content) to the activation of a cognitive structure in the *observer*. This focus has been adopted by other theorists, but there is considerable disagreement about the nature of this assumed cognitive structure. Nichols (2002), for example, doubts that a modular VIM as proposed by Blair (1995) would by itself be able to distinguish wrong from merely bad. Consider the example of a patient expressing agony while a nurse is changing the bandage of his wound (see Royzman et al., 2009). Although we instantly realize the patient's suffering and even that the nurse is its proximal cause, we are not inclined to morally condemn her. It is not clear, however, why the observer's VIM should not be activated in this example. It seems that we use some additional information when we make the moral/conventional distinction.

As already discussed in the section on "The Place of Emotions in Moral Judgments," Nichols (2002, 2004) argues that this additional information is contained in a person's "normative theory," which is acquired in a cultural learning process. The normative theory contains all socials rules, moral as well as conventional. Whether a given transgression elicits the signature moral pattern depends on whether the violated rule is backed by the activation of an emotion. This emotion need not be related to harm. For example, Nichols (2002) demonstrated that prototypically conventional transgressions can elicit the signature moral pattern if they are associated with disgust. His participants judged snorting loudly and spitting into cups at dinner tables to be universally and authority-independently wrong, especially those who scored high on a measure of disgust sensitivity (but see Royzman et al., 2009, for an alternative interpretation of the data).

Turiel's (1983) distinction of two distinct social domains that can be empirically identified through unique signature patterns has also been questioned. Kelly et al. (2007) argue that the elements of the signature moral pattern on the one hand, and those of the signature conventional pattern on the other, do not co-occur as monolithically as assumed by Turiel. For example, they point to the results of Haidt et al. (1993) and Nichols (2002) showing that people sometimes judge transgressions as universally wrong without justifying this assessment with notions of harm, fairness, or justice (see section on "Intuitionist Theories"). Thus, harm-related concerns do not seem *necessary* to elicit aspects of the signature moral pattern. Conversely, as Kelly et al. (2007) show, neither are they *sufficient* to do so. Instead of employing simple schoolyard transgressions of the kind Turiel (1983) and most of his followers used, Kelly and colleagues presented their participants with more complex harm-related cases from the adult world. For example, they asked their subjects whether it was okay for the captain of a modern U.S. cargo ship to whip one of his sailors as punishment for being drunk at work. Most participants responded that it was not. The researchers then went on to tell the same story, with the only difference that it was now supposed to have taken place several hundred years ago. The percentage of subjects considering the captain's whipping behavior as wrong decreased significantly in this scenario. Similar changes were obtained when norm violations were fictitiously placed in faraway countries. Similarly, Kelly et al. (2007) showed that whether harmful actions are judged as wrong often depends on whether these actions were approved or forbidden by an authority. Taken together, these findings indicate that intrinsically harmful actions are not necessarily seen as universally or authority-independently wrong, contrary to what was assumed by Turiel (1983) and other harm-based approaches (Royzman et al., 2009; Sousa et al., 2009).

In light of this evidence, Sripada and Stich (2006) argue that the distinction between conventional and moral rules is not psychologically meaningful. At the same time, however, they share the intuition that not all social rules are treated identically. Rather than contrasting moral rules with conventional ones, they argue that there is a psychologically important subclass of rules that they call "norms." Norms are not characterized by abstract philosophical principles or by specific contents, but mainly by the fact that people are *intrinsically motivated* to follow them.

In the Sripada and Stich (2006) model, people across all ages and cultures share the tendency to acquire and execute norms. The content of these norms, however, is assumed to be entirely determined by the social environment, and it need not (contrary to what Western philosophers have assumed) be held to be universally valid. This view has the advantage that it simultaneously explains the difference between norms and conventions, while making no assumptions about the contents of the norms a specific culture selects. However, to date the mechanisms implementing this assumed norm acquisition device have only been sketched, so it is hard to see how it can be empirically tested against theories that assume an innate preparedness for the acquisition of specific moral rules. Also, it is not clear how the content-free norm view can predict which of the culturally endorsed rules will be assigned the status of norms as opposed to conventions. Moreover, given the lack of constraints on content a larger diversity of norms might be predicted than is actually observed. If, however, commonalities are explained as solutions to similar adaptive challenges all social groups face, which would be a plausible claim, then it may be implausible to exclude evolutionary processes from creating some of these commonalities (see Haidt & Joseph, 2007; Joyce, 2006).

Moral Dilemmas

The currently most discussed and studied moral dilemma in both philosophy and psychology is the trolley dilemma, which we already have encountered in previous sections. Trolley dilemmas have become the Drosophila for testing alternative philosophical and psychological theories of moral judgments in harm-based moral dilemmas. This dilemma is theoretically interesting for philosophers because it can be shown that people seem to reason according to consequentialist principles in some versions of the dilemma, but according to deontological rules in other versions (Foot, 1967; Kamm, 2007; Thomson, 1985; Unger, 1996). In the past decade a large number of psychological studies have been performed to pinpoint the factors underlying the different moral intuitions.

An influential study based on 5,000 subjects in 120 countries was performed by Hauser et al. (2007). This study is the primary evidence for moral grammar theory (see section on "Moral Grammar Theory"). Figure 19.1 illustrates two basic trolley cases used by Hauser and colleagues. In their variant of the *bystander* dilemma, the driver of a train heading toward five people on the track faints. Denise, a passenger, has the option to redirect the train toward a side track with one person. Eighty-five percent of the subjects responded "yes" to the test question that asked whether it is permissible

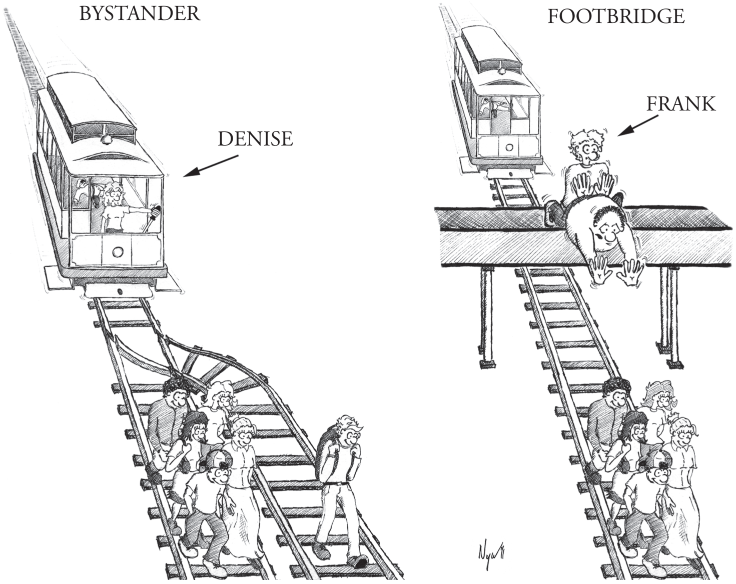

Fig. 19.1 Illustration of bystander and footbridge conditions (cf. Hauser et al., 2007).

for Denise to turn the train. In the *footbridge* version of the dilemma, a runaway train is also heading toward five people. Here Frank stands on a bridge going over the tracks, realizing that he can stop the train with a heavy weight. The only available heavy weight is a large man standing next to him. Just twelve percent respond "yes" to the test question of whether it is permissible for Frank to shove the man. The effect was reliably observed in the studied countries, although subjects generally had difficulties justifying their intuitions (see also Cushman et al., 2006). The authors interpret the effect as evidence for the unconscious use of the *doctrine of double effect* (DDE, see section on "Moral Grammar Theory"), which allows harming a person as a side effect, but not as a means of saving more people.

However, the two conditions differ in a number of potentially relevant features, giving rise to various alternative explanations. First, in one condition the agent is on the threatening train, and therefore part of the danger, whereas in the other condition the agent is not part of the dangerous situation. This difference might contribute to the effect. Second, in the footbridge condition the agent could potentially sacrifice himself instead of the person next to him (although the reference to heaviness may suggest that the other person will stop the train more efficiently). At any rate, this is not an option in the bystander dilemma. Third, in one condition the act involves a morally irrelevant object (the switch), whereas in the other condition the act involves forceful contact with the victim. Fourth, in one condition the intervention targets the threatening train, while in the other condition the victim is targeted. Fifth, the distance between the agent and the victim vary across conditions. Sixth, in one condition the potential alternative victim is only mentioned; in the other condition the victim is described as analogous to a heavy object. Seventh, the kind of death one imagines in the two conditions is more vivid and brutal in the footbridge than in the bystander condition; and eighth and last, the test questions differ in a way that can be expected to independently have an effect in the observed direction. Shoving a man is certainly considered less acceptable than turning a train even without the context of a trolley dilemma. A brief summary of the research of the past years is that it has been shown that almost all these confounding factors influence judgments, along with a number of others.

One plausible factor involves differences in *directness*. Previous research with other paradigms has shown that people find indirect harm less aversive than direct harm (Moore et al., 2008; Paharia et al., 2009; Royzman & Baron, 2002). Greene et al. (2009) have split this factor into three components: spatial proximity, physical contact, and personal force. Personal force refers to impacts on victims that are directly generated by muscular force of the agent. Touching and pushing a victim is an example of physical contact and personal force; using a pole for pushing would be an example of personal force without physical contact. Different versions of the footbridge dilemma were compared in which the three components were pitted against each other. The results show that moral permissibility assessments were explained by the personal force factor. In additional studies it was shown that only personal force that is due to an intentional act is relevant.

Personal force does not provide a full account of intuitions in trolley dilemmas, however. Waldmann and his colleagues (Waldmann & Dieterich, 2007; Waldmann & Wiegmann, 2010) have constructed versions of the trolley dilemma in which the victims in all conditions were sitting in vehicles (which made their kind of death comparable), the agents were remote, the acts were equated, and neither physical contact nor personal force was necessary to act. For example, in the bystander variant a train heading toward a train with five passengers could be redirected to a side track in which a train with one passenger is located by pressing a button in a remote control center. In the contrasted footbridge analog, the setup is the same but now the train on the side track with one passenger could be redirected by pressing a button onto the main track, where this train would stop the threatening runaway train and thus save the five (Waldmann & Wiegmann, 2010). Participants reliably found the intervention in the first condition in which the threatening train was redirected ("threat intervention") more acceptable than the one in the second condition in which the train with the one victim was redirected ("victim intervention"), although no personal force was involved.

One possible explanation of the differences between threat and victim intervention is the DDE (e.g., Cushman et al., 2010; Mikhail, 2011 Royzman & Baron, 2002). Whereas in the threat intervention condition the victim is harmed as a side effect of saving the five, in the contrasted victim intervention condition the victim is used as a means to stop the runaway train. The doctrine of double effect can only be supported by indirect evidence, not by asking subjects. Cushman et al. (2006) have shown that this rule is not consciously accessible to subjects who are requested to provide a justification for their moral judgment. Other moral rules, however, such as the principle that touching and thereby harming a person (contact principle) is impermissible, can be consciously accessed.

A popular paradigm to test whether using a person as a means to save others is really particularly aversive is based on Thomson's (1985) loop idea. In this variant of the bystander dilemma, the side track loops back to the main track right before the location where the five victims sit. This small variation turns the victim on the side track into a means to save the five. If the runaway train is redirected to the side track without being stopped by the person sitting there, it would go back to the main track and kill the five. Hauser et al. (2007) found that subjects judge the act in this condition more aversive than in the regular bystander (i.e., side effect) condition, but their experiment had the already mentioned confounds. Better-controlled studies did not find a difference (Greene et al., 2009; Waldmann & Dieterich, 2007; but see Sinnott-Armstrong, Mallon, McCoy, & Hull, 2008).

How else can the difference between threat and victim intervention be explained when personal force does not play a role in either condition? Waldmann and Wiegmann (2010) have proposed a *double causal contrast theory* to explain differences in intuitions in scenarios in which other relevant factors, such as distance, personal force, kind of victim, or kind of death have been held constant. The general idea motivating this theory is that people pay special attention to the intervention option when judging moral acceptability. Like all theories of moral judgments this theory predicts that reasoners are sensitive to the global contrast entailed by acting and nonacting (e.g., five victims vs. one), which explains why we differentiate between saving five or saving 1,000,000 (Bartels, 2008; Nichols & Mallon, 2006). However, whereas the DDE additionally is sensitive to the causal processes generated by the intervention (e.g., side effect vs. means), the double contrast theory assumes that we focus on the morally relevant target of intervention (i.e., threats or victims) and assess the harm directly caused by intervening on this target in contrast to the harm in which the target would be directly involved in the absence of the intervention. This local, counterfactual contrast focusing on the target of intervention will, according to this theory, heavily influence the acceptability rating.

How does the double causal contrast theory explain the two standard dilemmas? In the threat intervention condition, the proposed act can be summarized as redirecting the threat. Thus, the morally relevant target is the threatening trolley. To assess the local contrast, we need to focus on the direct harm caused by the target of intervention, which is one seriously harmed person. This outcome is contrasted with the direct harm caused by the target of intervention in the absence of the intervention, which in this condition are five harmed people. In contrast, in the victim intervention condition, the proposed act can be described as redirecting the victim in the train on the side track. Thus, the local contrast will focus on the train with its single potential victim. Setting this train into motion will directly cause harm to this victim. The fact that five people are saved further in the future is an indirect, more remote consequence of the act and therefore not part of the local contrast. The proposed intervention is contrasted with what would happen to the target of intervention in the absence of an intervention. In this case the person sitting in the train on the safe track would remain unharmed. The local contrast implies that the act is harmful, which predicts the lowered acceptability ratings.

Double causal contrast theory explains why people are not only sensitive to how the victim is directly harmed by the intervention but also that they consider whether the victim would have been harmed in the absence of an intervention (Moore et al., 2008). Moreover, Waldmann and Wiegmann (2010) showed that people accept harming a victim as a means when the local contrast is favorable. In one of their experiments they described a trolley dilemma in which the runaway train threatening five carries a passenger. If nothing was done, this passenger would stay alive and the train would kill the five. The five can be saved, however, if an empty train is redirected toward the threatening train, derailing it by pushing its passenger, who would die in the process, against the emergency brakes. Although here the intervention directly kills one person who plays the role of a means to save the five, subjects find this act highly acceptable. They focus on the threatening train with its single victim as the target of intervention, which leads to a contrast between five and one dead person.

The debate about the role of the DDE focuses on causal and intentional factors underlying moral intuitions. However, there are many other factors influencing judgments in trolley dilemmas. Rai and Holyoak (2010) demonstrate that domain-general factors that have been identified in behavioral economics also affect judgments in trolley dilemmas. Subjects who were asked to generate many reasons in favor of the action paradoxically rated it as *less* permissible than those who generated fewer reasons. This is consistent with research by Schwarz (1998), who showed that ease of retrieval of justifications is used as indicator for the quality of an option in nonmoral consumer choice. Other factors are mood of subjects (Strohminger, Lewis, & Meyer, 2011; Valdesolo & DeSteno, 2006), thinking styles (Bartels, 2008), preferred ethical position (Lombrozo, 2009), working memory (Moore et al., 2008), test question (see Kahane & Shackel, 2010), kind of victim (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009), vividness of death (Bartels, 2008), and the order of presenting different dilemmas (Iliev et al., 2009; Petrinovich & O'Neill, 1996; Wiegmann, Okan, & Nagel, in press). In short, it seems hopeless to look for the one and only explanation of moral intuitions in dilemmas. The research suggests that various moral and nonmoral factors interact in the generation of moral judgments about dilemmas.

The Role of Intention

A popular assumption implicit in many normative theories of morality is that we can only be held accountable for outcomes we have caused (Driver, 2008). We cannot possibly be responsible for bad events that are not directly or indirectly causally linked to our acts. In addition, most normative theories assign a special status to harm that was intentionally caused. For example, the doctrine of double effect (see sections on "Moral Grammar Theory" and "Moral Dilemmas") forbids intentionally harming a person, whereas unintentional harm may be permitted in some circumstances, even when the harmful outcome is foreseen. The interplay of intention and causal responsibility has also been central in descriptive theories of blame ascription (Alicke, 2000). A variety of recent studies have taken a closer look at the role of intentions and outcomes in moral judgments.

Cushman (2008) noted that different test questions may influence the relative contribution of these two components, intention and causation, in moral judgments. Cushman adopted a standard definition of intentional action according to which an act is intentionally performed if among other things the expected consequences are both desired and the act is believed to bring about these consequences (see Malle & Knobe, 1997). He used a story in which Jenny, the protagonist, is taking a course in sculpture and is assigned to work with a partner to weld together pieces of metal. The factors desire, belief, and consequences were manipulated independently using different cover stories: Jenny either desired or did not desire to burn her partner's hand, and she believed or did not believe that the act causes the harmful outcome. Moreover, it was varied whether the outcome did or did not occur. Cushman showed that judgments of blame and punishment are more sensitive to whether the harmful outcome caused by the agent occurs, whereas judgments of wrongness and permissibility are more sensitive to the agent's belief with respect to harming someone. However, the belief factor was the strongest predictor for both kinds of judgments.

Although our normative intuitions imply that we only should be held responsible for outcomes that are under our causal control, a number of experiments about *moral luck* (Williams, 1982) have shown that negative outcomes that are not fully under the agent's control may also influence our judgments. A typical example of moral luck is the following scenario: A father who bathes his child in a tub answers the phone in the nearby living room after telling his child to stay put. He believes that his son will indeed stay put so that nothing bad will happen. The father is typically judged to be more morally blameworthy if his child drowns (an unlucky outcome) than if his child stays safe (a lucky outcome). Thus, in both cases the intentions and the knowledge are the same, but the outcomes vary due to unforeseen random factors. In psychology, the apparently inappropriate weight given to the outcome has been labeled *outcome bias*, which has been documented in many different scenarios (see Baron & Hershey, 1988; Gino et al., 2010).

One obvious theory explaining outcome biases is that people give undue weight to the valence of the outcome, even though the agent did not intend it and is not fully responsible for it. However, Young, Nichols, and Saxe (2010) proposed an alternative theory according to which moral luck depends strongly on belief attribution and only indirectly on the bad outcome. The theory claims that the bad outcome provides evidence that the unlucky agent's beliefs are erroneous. Holding an erroneous belief that can cause harm is blameworthy and therefore leads to harsher moral judgments. In contrast, in the condition of the lucky agent the outcome validates the correctness of the agent's prior beliefs.

Another recent controversy revolves around the causal relationship between intentions and moral judgments. Cushman (2008) and Young et al. (2010) adopt the traditional assumption that this relationship is unidirectional: The agent's intention determines the moral judgment of an act. However, consider the following example presented by Knobe (2003):

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment." The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was harmed. (p. 191)

In a second version of this scenario, the word "harm" was replaced by "help." When subjects were asked whether they think the chairman intentionally harmed the environment, eighty-two percent answered in the affirmative. In contrast, in the help condition seventy-seven percent said that the agent did *not* bring about the good side effect intentionally. Knobe concluded that in judging whether the side effect was brought about intentionally, the moral value of the side effect is crucial. People seem considerably more willing to say that a side effect was brought about intentionally when they regard it as bad than when they regard it as good. This finding suggests the opposite from what is traditionally assumed: The moral evaluation of the outcome seems to determine whether intentionality is attributed, not the other way around.

There have been a lot of attempts to explain the so-called side-effect effect (also known as Knobe effect). Most of the proposed explanations can broadly be put into two groups (see Feltz, 2007; Uttich & Lombrozo, 2010). One group explains the side-effect effect by claiming that moral evaluations actually play a role in our concept of intentional action. According to this group of theories, the concept of intentional action actually includes and can be determined by the moral value of the effects that are caused by the act (see, e.g., Knobe, 2010; Mele & Cushman, 2007; Nichols & Ulatowski, 2007).

The other group denies this claim and explains the side-effect effect by arguing that subjects' judgments of intentional action are biased. For instance, Adams and Steadman (2004) invoke conversational pragmatics and argue that people want to express blame for the agent in the negative side-effect condition by characterizing the outcome as intentional. Another supporter of this view is Nadelhoffer (2004), who claims that subjects' judgments are biased because they get emotionally affected by the bad side effect. Guglielmo and Malle (2010) believe that task demands forced participants to use the term "intentional:" When given a choice, most participants prefer to say that the agent brought about the bad side effect *knowingly* rather than *intentionally* (similar to what the DDE would predict; see section on "Moral Grammar Theory").

Recently, Uttich and Lombrozo (2010) offered an interesting explanation that does not fall into either of these groups. They propose a rational explanation of the side-effect effect according to which the asymmetry in intentionality judgments arises because behavior that conforms to norms (moral or otherwise) is less informative regarding the mental state of the actor than norm-violating behavior (see Machery, 2008, for another theory of this kind). Prescriptive norms give us a positive reason to act, regardless of our intentions. Hence, a behavior that conforms to a norm does not tell us much about the agent's intentions. In contrast, violating a norm provides us with positive evidence about the agent's mental state. For example, in Knobe's examples the norm is not to harm the environment. Thus, when the chairman starts a program that helps the environment, we cannot tell whether he intends this or just follows a norm, whereas in the contrast case the norm-violating behavior provides us with strong evidence of an intention to harm the environment. This theory is not restricted to moral norms; rather, it applies to all norms, moral or nonmoral. Uttich and Lombrozo (2010) could show that virtually the same asymmetry can be observed when the cover stories mention the conventional norm that specific cars usually have a dark color. Knobe (2010) has recently offered a similar explanation, but in contrast to Uttich and Lombrozo he highlights the role of *moral* norms.

In sum, the present research indicates that the role of intention is far more complex than previously thought. Intentions are unobservable states that need to be inferred. Apparently, a large number of factors, including observed behavior, outcomes, causal structure, rationality assumptions, and norms, contribute to these attributions. Moreover, our language allows for subtle differentiations between different types of intentionality (e.g., desire, want, intend, foresee), which form a complex network with other factors underlying moral judgment.

An interesting direction for future research might be to take a closer look at the role of intentions in different moral domains and in different cultures. As for domain differences, Young and Saxe (2011) have shown that intentions are assigned more weight for moral judgments of harm violations, like assault, compared to purity violations, like incest. Thus, differences in the role of intentions between different cultures may arise due to differences in the culture-specific importance of moral domains. However, it is also possible that cultures differ within domains. In Western societies, the intentions of the agent are viewed as very important when assessing moral accountability, possibly more than in other cultures. Intentional transgressions of moral rules are typically condemned much more than accidental transgressions. In contrast, Rai and Fiske (2011) point out that in honor cultures, a woman who has sexual relations outside marriage, *even against her will*, defiles her family and is therefore punished.

Sacred/Protected Values

A characteristic feature of some moral values is that they resist trade-offs with other values. For example, many people find it impossible, inappropriate, or even outright abhorrent to put a price on human lives, friendships, democratic votes, or the preservation of the environment. It seems that some people ascribe infinite values to such entities, in that they would not accept any amount of any other good (especially not monetary ones) as compensation for the destruction or compromise of them. Such values have been termed "sacred" (Tetlock, Peterson, & Lerner, 1996) or "protected" (Baron & Spranca, 1997). Although both terms refer to the same phenomenon, the corresponding lines of research analyze it from different theoretical viewpoints, yielding different implications and even partially incompatible conclusions.

Tetlock and his colleagues describe sacred values (SVs) in their cultural context and analyze their social psychological functions. According to the revised Value Pluralism Model (Tetlock et al., 1996), people value different things for different reasons. When it comes to interpersonally relevant entities (such as intimate relationships, human rights, or religious symbols), people feel they have a commitment to others within their cultural community; they need to respect these entities in order to demonstrate that they are an estimable member of the community. The categorical nature of these commitments implies that within these sacred domains of social life, favors and goods are usually exchanged without numerical comparison (Fiske & Tetlock, 1997). If people compromise the respective values by trading them off against secular values (such as money, time, or convenience), they disqualify themselves from important social roles. Sacred and secular values are constitutively incommensurable; they cannot be sensibly compared, and mere attempts of comparison can destroy the SVs.

The motivation of people to hold SVs is to preserve their identity as full-fledged moral beings. If they witness others engaging in or merely contemplating taboo trade-offs, they typically react with moral outrage, a unitary response pattern consisting of harsh trait attributions, anger or contempt, and strong punitive impulses toward the offender (Tetlock, Kristel, Elson, Green, & Lerner, 2000). Due to unavoidable resource constraints in the real world, however, people are often forced to trade off SVs themselves. In such cases, they go to great lengths to conceal these trade-offs, for example, by means of decision avoidance or rhetorical obfuscation. Thus, people can be portrayed as both unapologetic defenders of SVs and at the same time as experts in finding ways to camouflage or overlook transgressions (Tetlock, 2003). Despite this discrepancy, Tetlock does not see people as hopeless hypocrites but instead as intuitive theologians striving to "[protect] sacred values from secular encroachments" (Tetlock, 2002, p. 452). Their rigidity is not seen as irrational but instead as serving important psychological functions and, on a larger scale, preventing subversion of meaningful cultural institutions (Fiske & Tetlock, 1997).

Baron approaches protected values (PVs) in the framework of the heuristics and biases program (see also section on "Moral Heuristics"). The main idea is that PVs are derived from deontological rules about acts (e.g., "do not kill"), irrespective of the consequences (Baron & Spranca, 1997). These rules are usually adaptive if treated as rules of thumb, but they may sometimes lead to suboptimal outcomes if they are unreflectively generalized to all contexts (Baron, 1998). In contrast to Tetlock, Baron reduces human values to a single utility metric, treating PVs as biases and stressing the problems they create for a utilitarian analysis.

One implication of the basis for PVs in absolute deontological rules is *quantity insensitivity*. For example, it seems to make only a small difference for people whether an act leads to greater or lesser harm to one of their PVs (Baron & Spranca, 1997), and some people seem to find it equally wrong to compromise a PV once or twice (Ritov & Baron, 1999). Another feature of deontological rules is that they usually prohibit harmful acts but not omissions, since prohibiting the latter would produce potentially unlimited obligations (Baron & Miller, 2000). Thus, PVs are seen as a source of omission bias (Baron & Ritov, 2009; Ritov & Baron, 1999) because actions are more likely to compromise PVs than omissions. For example, Ritov and Baron (1999) presented their subjects with a scenario in which 20 species of fish living in a river would become extinct unless a dam was opened. However, opening the dam would cause the extinction of two different species living downstream which would otherwise survive. People with a PV against extinguishing species were especially unwilling to open the dam, even though this decision would result in a greater net amount of damage to their cherished natural resource.

Many people seem to readily endorse statements implying PVs when asked directly (e.g., "This should be prohibited no matter how great the benefits from allowing it;" Baron & Spranca, 1997, p. 7). However, according to Baron and Leshner (2000), such judgments may be the result of reflexive, incomplete thinking which can be overcome quite easily. When PVs are challenged with realistic counterexamples involving extremely high benefits or low probabilities for harm to PVs, many people relativize their absolute claims. This finding is taken to indicate that expressions of PVs should not be taken too seriously.

This remarkable tension between rigidity and flexibility has recently been interpreted differently by Bartels, Medin, and colleagues. Instead of seeing deontological judgments as an impediment for consequentialist judgments, they regard both as often positively correlated across people (Iliev et al., 2009). That is, people holding PVs that rigidly prohibit certain actions in one task can be shown to be especially sensitive to consequences of these actions in different tasks, compared to people without PVs. Whether they give more weight to means or ends is largely a function of their attentional focus, which in turn is crucially affected by domain-general individual thinking styles, as well as low-level features of the task, such as framing and context effects (Bartels, 2008). For example, Bartels and Medin (2007) argued that the framing of the response options used by Ritov and Baron (1999) in the river diversion scenario ("Would you open the dam? Yes/No," followed by a measure for quantity sensitivity) directs the subjects' attention to the act of killing species. In this condition many people maintain a categorical prohibition against this act, which leads them to express a PV. Bartels and Medin (2007) went on to show that reframing the response alternatives so that they deflect attention away from the action itself to the consequences (by having subjects choose from a list of alternatives the maximum number of species living downstream they would be willing to kill by opening the dam to save the twenty species at risk) leads people with PVs to become *more* quantity sensitive, and less likely to show an omission bias than those without PVs. It seems as if the moral issue at stake is more central for people with PVs, and that they show amplified reactions in whatever direction their attention is steered by the task at hand (but see Baron & Ritov, 2009). In general, research on sacred and protected values provides an interesting test case showing that theories of moral judgments need to combine domain-specific cognitions (e.g., moral values) and domain-general mechanisms (e.g., attention).

Conclusions and Future Directions

The recent close cooperation both within psychology and across different disciplines has led to numerous new insights about morality. Summarizing the research from the viewpoint of a cognitive psychologist, three general research foci can be identified. First, many researchers have been interested in exploring the role of emotions and affects in moral judgments (see sections on "Emotion-based Theories" and "Dual-Process Theory"). This interest was initially motivated by a critique of previous paradigms (e.g., Kohlberg, 1981) in which conscious reasoning and rational deliberations were given a central place. In contrast, the more recent research has shown that many judgments are based on intuitions that are unconsciously elicited and are often accompanied by affects and emotions. The exact role of emotions is still not entirely clear. Emotions may precede or follow judgments, they may be constitutive for judgments, or they may be independent of rational judgment processes. A likely outcome of this debate may be that all of these possibilities occur, although we still need to know the boundary conditions of the different possibilities.

Second, the research on intuitions and emotions has largely addressed the global question of how reasoning and emotions in general are interrelated, but it has neglected the issue how specific intuitive judgments are caused. Thus, based on this research it is often impossible to make specific predictions about judgments for specific moral issues. The research has frequently been abstractly organized around a dichotomy between conscious reasoning and unconsciously elicited intuitions, which may have led to a neglect of research about the cognitive processes eliciting intuitions. In cognitive psychology, very few processes, not even logical reasoning and problem solving, are considered under full conscious control (see Evans, Chapter 8). Rather, cognitive theories specify the often unconscious information processing steps leading from an eliciting situation to a judgment. Although we still know little about these processes, some researchers have made progress in recent years specifying moral rules (e.g., doctrine of double effect) or moral heuristics underlying the appraisal of moral scenarios.

Third, an overarching question motivating most research on moral judgment is whether moral cognitions are special, or whether they represent just specific contents that otherwise can be handled by domain-general theories. The present research suggests that there is no innate specialized module devoted to morality that is encapsulated from other cognitive processes. Many studies that were motivated by domain-general theories, for example, behavioral economics, judgment and decision-making theories, or attention theories, have shown that moral reasoning is not an isolated process but rather recruits domain-general processes that may lead to phenomena also found in other domains. On the other hand, a full reduction of moral cognitions to general cognitions also seems implausible. Moral judgments use moral rules, moral values, or norms that have characteristics that differ from the general class of rules. They are typically accompanied by strong affect and emotions, which endow them with a force that goes beyond general conventional norms. Moral rules or norms are typically viewed as authority independent, as ends that have to be honored, as particularly important, and by some people as universally valid. Thus, there is a consensus in the literature that humans are born with dispositions to honor norms that manifest themselves in moral judgments. Whether beyond the general capacity to acquire norms, there is also an innate capacity that predisposes humans to acquire specific *moral* rules, is an open question that is currently strongly debated.

In this review, we concentrated on research about explicit judgment tasks. Some researchers have questioned whether studying isolated judgments, especially with controlled experimental tasks, is ecologically valid (Gigerenzer, 2010). Our position is that we should not primarily study moral judgments to predict behavior, but rather to understand how people judge what is right or wrong. People's opinions about moral issues, such as abortion, capital punishment, health, or food, are important factors shaping our society. However, it can be argued that implicit judgments are also reflected in *actions*. Although it is well known that moral judgments are not strongly correlated with corresponding actions, it is interesting to compare explicit with more implicit moral evaluations. There are several interesting lines of research investigating actual behaviors that can be viewed as indicators of implicit moral judgment. For example, it has been shown that people paradoxically feel licensed to behave in morally dubious ways (e.g., cheating, lying, not donating to charity, or making uncooperative decisions) when they have activated a particularly positive view of their moral self (Mazar, Amir, & Ariely, 2008; Mazar & Zhong, 2010; Sachdeva, Iliev, & Medin, 2009). Conversely, they feel compelled to act particularly morally when their moral self-image is threatened ("moral cleansing behavior," see Sachdeva et al., 2009; Tetlock et al., 2000), demonstrating the importance of self-regulatory processes for implicit judgments underlying moral behavior. People's implicit judgments concerning issues of fairness, altruism, cooperation, and punishment have also been assessed using behavioral measures. Fairness has been extensively investigated in simple bargaining games, primarily in the Ultimatum and Dictator Games, which investigate when subjects would reject unfair distributions of goods even when this implies that they would not get anything (e.g., Camerer, 2003; Camerer & Smith, Chapter 18). Common good games, which study individuals competing with other members of a group for common resources, have been used to obtain behavioral measures of cooperation, defection, and punishment (e.g., Fehr & Gächter, 2000). There is also a huge literature on altruism and prosocial behavior (see Batson, 2011).

We have seen that, although there seems to be an explosion of research on morality in recent years, many questions remain unanswered. Here we just list a few of these questions that seem particularly pressing from the viewpoint of cognitive psychology. For example, we know very little about the appraisal processes leading to moral judgments. Most of the theories dealing with appraisal have been developed in the context of very limited paradigms (e.g., trolley problems), so that the generality of these theories is unknown. Moreover, oversimplified theories of the representation of moral norms have postulated rules that seem to only superficially fit the investigated task. "Do no harm," for example, is certainly a rule that often seems plausible, but it does not capture the context sensitivity that people's judgments display. Thus, if a rule-based account is chosen, a much more complex system of rules needs to be specified, which includes boundary conditions and exceptions. Moreover, if research on categorization is taken as a model (see Rips et al., Chapter 11), we need to ask whether rules are the only plausible format for the representation of moral knowledge or whether other representational devices, such as exemplars, prototypes, schemas, or analogies, also play a role.

We expect more research concerning the interplay of domain-general and domain-specific process in moral judgments. As the research on trolley dilemmas (see section on "Moral Dilemmas") shows, it seems necessary to negotiate the relative role of these processes for each target problem separately. There has been a tendency in the field to overstate findings as evidence for the use of grand philosophical positions. In our view, it seems implausible to argue that a sociopath reasons like a consequentialist when a much simpler account can be found. The fact that somebody finds smothering a baby abhorrent, or that somebody finds it preferable that one person instead of 1,000,000 people dies, does not turn this person into a deontologist or a consequentialist. It seems more plausible to pinpoint the reason for different judgments on more local factors, such as selective attention to specific aspects of a situation or deficits of affective processing.

Our review was largely limited to studies focusing on Western moral norms (e.g., prohibition of harm), which have been central in studies on the cognitive and affective foundations of moral judgment. The explanation for this one-sidedness is that both researchers and research subjects typically have a Western background (Henrich, Heine, & Norenzayan, 2010). Although anthropology has collected massive evidence showing that there is more to morality than concerns about harm or fairness/justice, most of this research so far is descriptive. We know that other cultures often endorse other norms, but we do not know how moral cognitions in other societies differ from ours. Do people in other cultures employ the same cognitive processes but invoke different moral rules, or are the cognitive processes underlying judgments different in other cultures? The most likely answer is that both possibilities may turn out to be true. If specified very abstractly, a process such as attentional focus will certainly influence judgments in different cultures, although the target of the focus will of course shift. On the other hand, we do not know whether general regularities that go beyond specific rules but are less abstract than attention universally play a similar role. For example, the section on "The Role of Intention" highlighted the role of intention in moral blame. An interesting question might be whether intentional attributions and the weighing of intentions are similar in different domains and in different cultures. In sum, moral cognitions are most certainly an interesting topic for future research, but we have only started to understand this fascinating competency.

Acknowledgments

We thank K. Holyoak, T. Rai, and H. Rakoczy for helpful comments.

References

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, *64*, 173–181.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*, 1–42.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. New York: Oxford University Press.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values? *Journal of Experimental Psychology: Applied*, *6*, 183–194.

Baron, J., & Miller, J. G. (2000). Limiting the scope of moral obligations to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology*, *31*, 703–725.

Baron, J., & Ritov, I. (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 133–167). San Diego, CA: Elsevier.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*, 1–16.

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

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*, 24–28.

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

Batson, C. D. (2011). *Altruism in humans*. Oxford: Oxford University Press.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*, 187–202.

Binmore, K. (2008). *Rational decisions*. Princeton, NJ: Princeton University Press.

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

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision making*, *3*, 121–139.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Caruso, E. M., & Gino, F. (2011). Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. *Cognition*, *118*, 280–285.

Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.

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

Crain, W. C. (1985). *Theories of development*. Upper Saddle River, NJ: Prentice-Hall.

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

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

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 47–71). Oxford, England: Oxford University Press.

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

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Driver, J. (2008). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 423–440). Cambridge, MA: MIT Press.

Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences*, *11*, 373–378.

Dwyer, S. (2006). How good is the linguistic analogy? In P. Carruthers, S. Lawrence, & S. Stich (Eds.), *The innate mind: Culture and cognition* (pp. 237–256). New York: Oxford University Press.

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science*, *22*, 295–299.

Evans, J. St. B. T. (2007). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology*, *44*, 1204–1209.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior*, *28*, 265.

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, *18*, 342–350.

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology*, *18*, 255–297.

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

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded. *Topics in Cognitive Science*, *2*, 528–554.

Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

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

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*, 93–101.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*, 1029–1046.

Greene, J. D. (2008). The secret joke of Kant' s soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality* (pp. 35–79). Cambridge, MA: MIT Press.

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

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

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

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

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*, 1635–1647.

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., & Hersh, M. A. (2001). Sexual morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology*, *31*, 191–221.

Haidt, J., & Joseph, C. (2007). The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 367–391). New York: Oxford University Press.

Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 797–832). Hoboken, NJ: Wiley.

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

Harman, G., Mason, K., & Sinnott-Armstrong, W. (2010). Moral reasoning. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 206–245). Oxford, England: Oxford University Press.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: HarperCollins.

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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*, 1–6.

Hume, D. (1960). *An enquiry concerning the principles of morals*. La Salle, IL: Open Court. (Original work published in 1777).

Iliev, R., Sachdeva, S., Bartels, D. M., Joseph, C., Suzuki, S., & Medin, D. L. (2009). Attending to moral values. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin, (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 169–192). San Diego, CA: Elsevier.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*, 521–537.

Joyce, R. (2006). *The evolution of morality*. Cambridge, MA: MIT Press.

Kahane, G., & Shackel, N. (2008). Do abnormal responses show utilitarian bias? *Nature*, *452*, 5.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgment. *Mind and Language*, *25*, 561–582.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (in press). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–294). Cambridge, England: Cambridge University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kamm, F. M. (2007). *Intricate ethics*. Oxford, England: Oxford University Press.

Kant, I. (1959). *Foundation of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis, IN: Bobbs-Merrill. (Original work published in 1785).

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, affect, and the moral/conventional distinction. *Mind and Language*, *22*, 117–131.

Kern, M. C., & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science*, *20*, 378–384.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–194.

Knobe, J. (2006). The concept of intentional action: A case study in uses of folk psychology. *Philosophical Studies*, *130*, 203–231.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329.

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

Kohlberg, L. (1981). *The philosophy of moral development*. San Francisco, CA: Harper.

Laham, S. M., Alter, A. L., & Goodwin, G. P. (2009). Easy on the mind, easy on the wrongdoer: Discrepantly fluent violations are deemed less morally wrong. *Cognition*, *112*, 462–466.

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, *46*, 352–367.

Litman, L., & Reber, A. S. (2005). Implicit cognition and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 431–453). New York: Cambridge University Press.

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

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, *23*, 165.

Machery, E., & Mallon, M. (2010). The evolution of morality. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 3–46). Oxford, England: Oxford University Press.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*, 633–644.

Mazar, N., & Zhong, C. (2010). Do green products make us better people? *Psychological Science*, *21*, 494–498.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*, 577–580.

Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, *31*, 184–201.

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

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

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions, and the utilitarian brain. *Trends in Cognitive Sciences*, *11*, 319–321.

Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: A working model. *Neuroreport*, *14*, 299–305.

Moll, J., de Oliveriera-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, 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*, 2370–2736.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*, 549–557.

Moore, A. B., Stevens, J., & Conway, A. R. (2011). Individual differences in sensitivity to reward and punishment predict moral judgment. *Personality and Individual Differences*, *50*, 621–625.

Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*, *24*, 196.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*, 221–236.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*, 530–542.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, *22*, 346–365.

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, *109*, 134–141.

Parfit, D. (2011). *On what matters*. Oxford, England: Oxford University Press.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, *2*, 511–527.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, *17*, 145–171.

Piaget, J. (1932). *The moral judgment of the child*. London: Kegan Paul, Trench, Trubner and Co.

Prinz, J. J. (2007). *The emotional construction of morals*. Oxford, England: Oxford University Press.

Prinz, J. J. (2008). Acquired moral truths. *Philosophy and Phenomenological Research*, *77*, 219–227.

Prinz, J. J., & Nichols, S. (2010). Moral emotions. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 111–146). Oxford, England: Oxford University Press.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*, 57–75.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science*, *34*, 311–321.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Reyna, V. F., & Casillas, W. (2009). Development and dual processes in moral reasoning: A fuzzy-trace theory approach. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making*. *The psychology of learning and motivation: Advances in research and theory* (pp. 207–236). San Diego, CA: Elsevier.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes*, *79*, 79–94.

Roedder, E., & Harman, G. (2010). Linguistics and moral theory. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 273–296). Oxford, England: Oxford University Press.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*, 165–184.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, *112*, 159–174.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*, 523–528.

Sachdeva, S., & Medin, D. L. (2008). Is it more wrong to care less? The effects of "more" and "less" on the quantity (in) sensitivity of protected values. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1239–1243). Austin, TX: Cognitive Science Society.

Scanlon, T. M. (1999). *What we owe to each other*. Cambridge, MA: Harvard University Press.

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

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review*, *2*, 87–99.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997), The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.

Simpson, E. (1974). Moral development research: A case study of scientific cultural bias. *Human Development*, *17*, 81–106.

Singer, P. (1979). *Practical ethics*. Cambridge, England: Cambridge University Press.

Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 2: The cognitive science of morality: Intuition and diversity* (pp. 47–76). Cambridge, MA: MIT Press.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind and Language*, *23*, 90–106.

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris & The Moral Psychology Research Group (Eds.), *The moral psychology handbook* (pp. 246–272). Oxford, England: Oxford University Press.

Sloman S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 1–26). San Diego, CA: Elsevier.

Sousa, P., Holbrook, C., & Piazza, J. (2009). The morality of harm. *Cognition*, *113*, 80–92.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*, 76–105.

Sripada, C. S. & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 2, pp. 280–301). New York: Oxford University Press.

Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition*, *119*, 295–300.

Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*, 531–573.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*, 451–471.

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

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*, 853–870.

Tetlock, P. E., Peterson, R. S., & Lerner, J. S. (1996). Revising the value pluralism model: Incorporating social content and context postulates. In C. Seligman, J. Olson, & M. Zanna (Eds.), *Ontario Symposium on Social and Personality Psychology: Values* (pp. 25–51). Mahwah, NJ: Erlbaum.

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

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

Turiel, E. (2006). The development of morality. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology. Vol. 3: Social, emotional, and personality development* (pp. 789–857), Hoboken, NJ: Wiley.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*, 479–491.

Unger, P. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*, 87–100.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*, 476–477.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*, 247–253.

Waldmann, M. R., & Wiegmann, A. (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In S. Ohlsson&R.Catrambone(Eds.),*Proceedingsofthe32ndAnnual Conference of the Cognitive Science Society* (pp. 2589–2594). Austin, TX: Cognitive Science Society.

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

Wiegmann, A., Okan, Y., & Nagel, J.(in press). Order effects in moral judgment. *Philosophical Psychology*.

Williams, B. (1982). Moral luck. In *Moral luck. Philosophical papers 1973–1980* (pp. 20–39). Cambridge, England: Cambridge University Press.

Williams, B. (1985). *Ethics and the limits of philosophy*. London: Fontana.

Woodward, P. A. (2001). *The doctrine of double effect*. Notre Dame, IN: Notre Dame University Press.

Wright, J. C., & Baril, G. (2011). The role of cognitive resources in determining our moral intuitions: Are we all liberals at heart? *Journal of Experimental Social Psychology*, *47*, 1007–1012.

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*, 333–349.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*, 149–298.