12.3 The Psychology and Rationality of Moral Judgment

Alex Wiegmann and Hanno Sauer

Summary

The field of moral psychology has become increasingly popular in recent years. This chapter focuses on two interrelated questions: first, how do people make moral judgments? It will review the most prominent theories in moral psychology that aim to characterize, explain, and predict people's moral judgments. Second, how should people's moral judgments be evaluated in terms of their rationality? This question is approached by reviewing the debate on the rationality of moral judgments and moral intuitions, which has been strongly influenced by findings in moral psychology but also by recent advances in learning theory.

1. Rationality and Morality

The rationality of moral judgment is one of the most contested issues in moral psychology. In this chapter, we will outline the cutting edge of contemporary work in moral psychology, convey a sense of the most promising recent developments, and explain where this research is currently heading. We will also address some of the philosophical implications these theories are supposed to yield: Are our moral intuitions robust and trustworthy? Or are they frail and unreliable? Ultimately, these *normative* questions regarding the rationality of moral judgment are what most philosophers and probably also many psychologists are interested in.

In the first part, we will summarize theories that make specific predictions about which moral judgments people are likely to make in response to which types of cases and what the psychological processes are that explain the patterns so discovered. In particular, we will explain how a theory based on the distinction between *model-free* and *model-based* information processing and decision making has come out as the main contender among such predictive theories. In doing so, we will integrate both insights from computational approaches to reinforcement learning as well as recent developments in dual-process models of cognition. This concludes the first, *descriptive* part of our overview.

The second, *normative* part addresses the issue of the rationality of moral judgments head-on.[1] The recent debate suggests that this question might not be best addressed in terms of the psychological foundations of moral cognition: Are moral judgments intuitive or controlled? Are they based on emotion or reason? Does reasoning produce, or merely rationalize, moral judgments? Rather, addressing the question of whether the processes generating moral beliefs are capable of incorporating and updating on morally salient information, that is, capable of *learning*, seems the most promising line of research. We will explain why some authors argue that the sophistication of processes of moral learning provides reasons for optimism regarding the rationality of moral judgments and why some authors continue to strike a more pessimistic note.

2. Predictive Theories of Moral Judgment

In this descriptive part of the chapter, we will review three prominent predictive theories of moral judgments, namely, Greene's dual-process model, Mikhail's universal moral grammar theory, and Cushman's and Crockett's model-free versus model-based approach, with a focus on the latter since it is the most recent and probably the most promising theory. It should be noted that the distinction between "descriptive" and "normative" accounts of moral cognition is at least somewhat artificial, as some of the models discussed in the former also have bearing on the latter issue. What we mean to emphasize here is that some theories in moral psychology make specific, testable predictions regarding how subjects will actually respond to certain issues and why, while the latter focus on deriving possible normative conclusions from an overall assessment of the state of the art in the psychology of moral judgment.

2.1 Dual-Process Models

The contemporary debate on the rationality of moral judgment is dominated by *dual-process* models of moral cognition (Greene, 2013). According to such models, cognitive operations are carried out by two fundamentally different types of processing, one quick, intuitive, automatic, evolutionarily old, and frequently unconscious (System 1), the other slow, effortful, analytic, evolutionarily young, and consciously controlled (System 2; Evans & Stanovich, 2013; Kahneman, 2011).

Some of these theories hold that moral judgment is System 1 all the way down (see section 3.1). Others have argued that the two types of processing can be mapped onto two different subsets of moral judgment: deontological and consequentialist moral judgments (Greene, 2008, 2013, 2014). This distinction is imported from philosophical moral theory and refers, roughly, to the views that the outcome of an action alone determines its moral status (consequentialism) or that other factors—such as a person's intentions or the way an outcome was brought about (e.g., as a means or a foreseen side effect)—matter as well (May, 2014b). This alternative is typically operationalized by classifying subjects' responses to certain well-known moral problems ("trolley cases"; for an overview, see Waldmann, Nagel, & Wiegmann, 2012), in which participants have to imagine a train about to run over five people and are asked whether it would be OK to perform a certain action in order to save the lives of the five people. The two most prominent variants of such trolley cases are usually referred to as *Switch* and *Push*, respectively. In Switch, one can, by flipping a switch, divert the train onto a different track, where it would run over and kill a single person. In Push, the train can be stopped by pushing a large man, who is standing on a bridge crossing the tracks, in front of it. While interventions in these cases (flipping the switch and pushing the man, respectively) are classified as *consequentialist* responses, not performing the action necessary to save the five people is usually considered the *deontological* option.[2]

Some want dual-process accounts to *debunk* deontological theories and *vindicate* consequentialist ones. The former are purportedly based on crude, alarm-like emotions (see also the moral-heuristics approach by Gigerenzer, 2008, or Sunstein, 2005) and the latter on flexible, sophisticated cost–benefit analyses (Greene, 2008). This pattern seemed to be reflected by patterns of brain activation, differences under cognitive load, and greater response times for consequentialist judgments, indicating controlled intuition override (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Moreover, the etiology of deontological intuitions seemed to show that they pick up on morally irrelevant differences such as how "up close and personal" a person is harmed (Greene et al., 2001).

Criticism of the dual-process model has focused on four issues: first, that the empirical data from neuroimaging (Machery, 2014; Poldrack, 2006; Sauer, 2012b) or response times (McGuire, Langdon, Coltheart, & Mackenzie, 2009) are more ambiguous than originally thought. Second, some have argued that the empirical data from neuroscience or certain evolutionary considerations play no actual role in the argument against deontology (Berker, 2009; cf. Greene's [2010] response). Third, it's not obvious why the emotional or intuitive basis of deontological judgments should pose any epistemic problem for them at all (Railton, 2014). Fourth, and most important, many have argued that even *if* the empirical data are sound, and even *if* they do the moral heavy lifting, and even *if* they do identify a rationally defective process, it remains unclear whether they really target deontological theories. The reason for this doubt is that the link between consequentialist moral judgment and controlled cognition can be reversed when the consequentialist option is the intuitive one and the deontological option requires override (Kahane et al., 2011). Moreover, in order for people's moral judgments to count as genuinely consequentialist, they would have to stem from an impartial concern for the greater good. This seems not to be the case (Kahane, Everett, Earp, Farias, & Savulescu, 2015). In general, the Trolleyological paradigm in moral psychology tends to misconstrue consequentialism (Kahane et al., 2018; but see also Conway, Goldstein-Greenwood, Polacek, & Greene, 2018; see also endnote 2).

2.2 Universal Moral Grammar

The dual-process model assumes that moral judgments result from domain-general processes that are not specific to morality. A further possibility is that people are endowed with an innate moral faculty that allows them to cognitively structure actions in terms of moral categories and to assign deontic statuses (right or wrong, permissible or impermissible) to those actions depending on this structure. Drawing on ideas pioneered in a linguistic context, various authors have tried to show that moral cognition is based on such a *universal moral grammar* (Hauser, Young, & Cushman, 2008; Mikhail, 2007, 2011; cf. Prinz, 2008).

People start to appreciate certain moral distinctions at a very early age, many of those distinctions seem

to be culture-invariant, and people have a hard time articulating the rules structuring their moral thinking (Hauser, Cushman, Young, Jin, & Mikhail, 2007). Also, some patterns in people's moral judgment are difficult to account for from within a dual-process framework, such as the distinction between harming as a means and harming as a foreseen side effect (e.g., Switch versus TrapDoor; see next section). Primarily, however, the nativist case for a universal grammar is supported by so-called poverty-of-the-stimulus arguments, according to which the moral rules that implicitly inform people's moral judgments are too subtle to be read off the scarce available evidence; subjects *couldn't* learn these rules empirically. We will return to this question after a short detour.

2.3 Model-Free versus Model-Based Moral Judgment

Crockett (2013) and Cushman (2013) independently developed a dual-system account of moral psychology that has already had an impact on the field of moral psychology and will likely continue to do so (e.g., Greene, 2017; Railton, 2017). According to this approach, two distinct systems evaluate actions: while the model-based system selects actions based on inferences about its (expected) outcomes in a specific situation, the model-free system assigns value to actions intrinsically, based on past experience (Crockett, 2013; Cushman, 2013).

Such dual-system frameworks have been argued to be of great value for studying and understanding judgment and decision making in general (Evans & Stanovich, 2013; Kahnemann, 2011; see Kruglanski & Gigerenzer, 2011, for a critical assessment), and Greene's emotional-versus-cognitive framework has been very influential in the field of moral psychology (e.g., Greene et al., 2001). However, it has been argued that this particular distinction of processes (i.e., emotional versus cognitive) might not be fully adequate (Cushman, 2013; Pessoa, 2008) and computationally too crude (Crockett, 2013; Mikhail, 2007; see Crockett, 2016, for the advantages of formal models) to characterize and explain moral judgments. Crockett and Cushman's distinction of action evaluation versus outcome evaluation is supposed to avoid the shortcomings of Greene's characterization of the two systems while keeping the benefits of a dual-system framework (Greene [2017] seems to agree). Findings from studies dissociating the aversiveness of actions from outcomes provide some initial evidence for separate action- and outcome-based (model-free versus model-based) value representations (Cushman, Gray, Gaffey, & Mendes, 2012; Miller & Cushman, 2013; Miller, Hannikainen, & Cushman, 2014).

The distinction between action- and outcome-based value representations resembles a distinction from computational approaches to reinforcement learning (cf. Cushman, 2013), and recent advances in neuroscience have identified corresponding neural signatures (Dayan & Niv, 2008; Dolan & Dayan, 2013; Gläscher, Daw, Dayan, & O'Doherty, 2010; Wunderlich, Dayan, & Dolan, 2012). Reinforcement learning investigates how agents (humans, rats, algorithms) can learn to predict the consequences of and optimize their actions in interaction with the environment (Dayan & Niv, 2008). Learning algorithms can be classified broadly as model free or model based (Sutton, 1988; Sutton & Barto, 1999, 2018), depending on how they learn and select actions in a given situation (for an example in human behavior, see Dayan & Niv, 2008). Put simply, a model-based algorithm builds a causal representation of its environment, uses this model to consider the whole range of effects that different actions would have, and chooses the action with the highest expected value. In contrast, a model-free algorithm does not rely on a model of the environment and thus cannot make farsighted choices. In a given situation, it simply selects the action with the best track record, which can be acquired by rather simple means such as prediction error (Gläscher et al., 2010) and temporal difference learning (Sutton, 1988; cf. Cushman, 2013). Although a model-free algorithm is computationally cheap, it can often lead to successful behavior. However, if the environment changes, it cannot adapt to these changes as readily as the computationally rather expensive model-based algorithms since it needs to gather experience in the new environment to update the track record accordingly (Crockett, 2013; Cushman, 2013; Greene, 2017).

Can the model-free-versus-model-based (or action-versus-outcome) framework explain prominent findings in moral psychology? Let us start with applying the framework to three variants of the well-known trolley dilemma: Switch, Push, and TrapDoor (the latter being similar to Push, but rather than the heavy person being pushed from the bridge, the flip of a switch opens a trap door through which the large man falls onto the tracks). The relation among the approval rates for the respective actions in these dilemmas is Switch > TrapDoor > Push (Mikhail, 2007, 2011). The model-free-versus-model-based framework can provide a straightforward explanation for the fact that Push is usually evaluated worse than the other two cases: although the model-based evaluation of the actions does not differ (since they lead to the same outcome), pushing a person usually leads to negative experiences (punishment, distress cues), whereas flipping a switch is not associated with a negative action

value. In other words, the model-free evaluation prefers flipping a switch to pushing a person (Crockett, 2013; Cushman, 2013). Things get more complicated when comparing the Switch and the TrapDoor variant. Here, not only the outcome (saving five lives) but also the required action (flipping a switch) is the same. What differentiates these two cases is the causal role of the victim. In Switch, the one person getting hit by the train is not causally necessary for saving the five lives (imagine the one person jumping off the track before getting hit), whereas in TrapDoor, the five people could not be saved without making use of the heavy person as a trolley stopper (Mikhail, 2007; Wiegmann & Waldmann, 2014). The model-free-versus-model-based framework attempts to capture this difference by postulating that the model-based system represents harming a person as a subgoal in TrapDoor but not in Switch. Subgoals, in turn, can be treated by the model-free system as if they were actions—and if the subgoal is aversive, as it is in the case of a person getting hit by a train, the model-free system assigns a negative value to it.[3] It is this negative value assignment of the model-free system that differentiates side effect cases from means cases and can explain why the latter are often judged as more aversive. A further popular finding in moral psychology is that usually people consider harmful actions worse than omissions with identical effects (Baron & Ritov, 2004). This asymmetry can be explained by pointing out that model-free values are only assigned to actions but not to the always available option of *not* performing an action.

To sum up, the model-free-versus-model-based account by Crockett (2013) and Cushman (2013) offers a promising domain-general theory of moral judgment.

3. The Rationality of Moral Judgment

In the first part of this chapter, we described prominent theories, findings, and developments in moral psychology. Not surprisingly, such insights came along with discussions about how to evaluate people's moral judgments and moral intuitions. In this second part of the chapter, we will address this normative question by first retracing the debate on the rationality of moral judgments from the beginning of moral psychology to the current state of the field on a rather abstract level. We then narrow down the considered time frame and content by focusing on the rationality of moral intuitions and how this debate has evolved in recent years due to advances in learning theory.

3.1 Classic Rationalism, Social Intuitionism, and Sentimentalism

Classical rationalism According to what has been the dominant paradigm in moral psychology for the second half of the 20th century, the rationality of moral judgment—or lack thereof—has to be sought in how the capacity for moral judgment and reasoning *develops*. Moral reasoning becomes increasingly sophisticated with ontogenetic stages. While infants may start out with a *preconventional* moral code in terms of external norms and sanctions, older children construe morality as a *conventional* matter of playing one's role and sticking to the rules; with adolescence, some people come to see morality as a postconventional affair regulated by universal principles (Kohlberg, 1969).

On the other hand, one of the hallmarks of moral judgment—such as the distinction between moral and conventional norms—is already appreciated at a very early age (Smetana & Braeges, 1990; Turiel, 1983), and the Kohlbergian account implicitly operates within a male perspective, thereby dismissing moral outlooks that are focused on maintaining a complex web of social interactions between concrete individuals, rather than on preserving abstract rights, principles of justice, or procedures of deliberation, as somehow undeveloped and immature (Gilligan, 1982).

The antirationalist turn With the beginning of the 21st century, rationalist dominance began to erode. The social intuitionist model, for instance, holds that moral judgment is a largely intuitive and automatic enterprise; reasons are typically produced post hoc, if at all; and when they are, it is not in the unmotivated pursuit of truth but for the purpose of social persuasion (Haidt, 2001; Haidt & Björklund, 2008). When people's moral judgments are debunked, they defend rather than question them. When this turns out to be unsuccessful, they don't abandon their beliefs but enter a state of being *morally dumbfounded*. Moral judgment, it now seemed, is not just largely intuitive, but which intuitions people have is due to differences in their *moral foundations*, with important political consequences (Graham, Haidt, & Nosek, 2009; Iyer, Koleva, Graham, Ditto, & Haidt, 2012; cf. Curry, 2016; Sauer, 2015).

Sentimentalism Others are less concerned with the proximal processes that yield particular moral judgments than with the psychological foundations of valuing itself. According to *sentimentalism* about moral judgment, values are distally grounded in (dispositions for) emotional reactions (Nichols, 2004; Nichols, Kumar, Lopez, Ayars, & Chan, 2016; Prinz, 2006, 2007, 2016). The empirical evidence suggests that moral judgments are constituted

by emotions, because emotions are both necessary and sufficient for moral judgment. Conditions such as psychopathy suggest that without the appropriate affective reactions, moral judgment is impaired (Blair, 1995). A battery of studies suggests that changing someone's emotions affects their moral beliefs. Dispassionate moralizing seems to be nonexistent. On the other hand, both the interpretation of this evidence (Sauer, 2012a) and its robustness have been called into question. The effect of (incidental) affect on moral judgment seems to be meagre (Landy & Goodwin, 2015; May, 2014a). Psychopaths do master the moral–conventional distinction after all, despite their affective impairments.

Rationalist replies Partly due to these developments, rationalists about moral judgment have started to speak up again. For some, the issue is partly conceptual: moral judgments must in some way be responsive to reasons; otherwise, they don't qualify as genuinely moral (Kennett & Fine, 2009). Sentimentalists are also accused of misidentifying their subject matter. What they have studied are time-slice responses to far-fetched toy problems (Gerrans & Kennett, 2010) rather than people's temporally extended moral agency. Others emphasize that rational processes can "migrate" into intuitive ones (Sauer, 2017). Still others have offered hybrid theories according to which moral judgments are compound cognitive–motivational states in which belief and emotion homeostatically cluster together (Kumar, 2016).

A third interesting development is due to research on how moral judgment affects various *other*, seemingly nonmoral, cognitive domains. Many studies conducted in the so-called experimental philosophy (Kauppinen, 2007) paradigm, for instance, focus on how causal or social cognition lies downstream from moral cognition: the moral, evaluative, or otherwise normative beliefs people hold influence their judgments about what causes what (Knobe & Fraser, 2008), who did what intentionally or not (Knobe, 2003), or what someone's personal identity is (Newman, De Freitas, & Knobe, 2015). This debate, too, has led to interesting discussions regarding the legitimacy of such an influence (Sauer & Bates, 2013) and whether it makes sense for us to be "moralizers through and through" (Knobe, 2010).

3.2 Moral Intuitions and Rational Learning

Recently, the debate on the rationality of moral judgments has focused on moral intuitions. Do our moral intuitions track morally relevant features in a reasonable way, suggesting that we should grant them high evidentiary status and be optimistic about their ability to successfully navigate us through, and solve problems in, the social world? Or does the relevant evidence advise us to paint a darker picture? In this section, we trace how these discussions have evolved in recent times due to advances in understanding the origins of our intuitions in general and of our moral intuitions in particular. For what follows, we closely follow Haidt's (2001, p. 818) characterization of the subject matter, according to which moral intuition is the sudden appearance in consciousness of a moral assessment, 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—and these moral intuitions are assumed to play a crucial role for moral judgments and decisions, although they can be overridden by other considerations.

Especially with the publication of the seminal articles by Haidt (2001) and Greene and colleagues (2001) at the beginning of the 21st century, the hitherto dominant rationalist view of moral judgments (Kohlberg, 1969; Piaget, 1932; Turiel, 1983) was seriously called into question by highlighting the role of moral intuitions, which were characterized as a rather bad guide for moral judgments and decision making. Moreover, this view seemed to be well supported by decades of psychological research indicating that intuitions in general are often misleading (Kahneman, 2003, 2011).

Relatedly but separately, there are interesting recent developments in philosophical moral epistemology that can both learn from and inform current psychological research. In particular, philosophers are keen to explore the issue of whether we can trust our moral intuitions and under what conditions and whether doing so requires a robust notion of mind-independent moral truth. Many philosophers argue that, despite what we know about the distal and proximal origins of our moral beliefs, we need not abandon most, or at least only some, of them (Bengson, Cuneo, & Shafer-Landau, 2019; Enoch, 2011; Huemer, 2008).

Recent empirical accounts of intuitive cognition also seem to somewhat vindicate this philosophical optimism regarding the rationality of moral judgment. Due to advances in research on the computational underpinnings of learning, especially in reinforcement learning (e.g., Mnih et al., 2015) and Bayesian learning (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011), it does not seem unreasonable to assume that our intuitions are the output of highly sophisticated learning systems that are rather flexible and domain general (cf. Railton, 2017). The findings suggest that our intuitions

can be finely attuned to even subtle contours of the decision landscape and often align with normative models of judgment and decision making such as, for example, Bayesian updating, although we often do not have deliberative access to these learning processes.

These advances in understanding the origins of our intuitions have been used by several authors to argue that the rather pessimistic assessment of moral intuitions may have to be revised as well. Perhaps most prominently, Railton (2014, 2017) argues that some popular findings in moral psychology that have been used to paint people's moral intuition in an irrational light should be reinterpreted. For instance, it was argued that people enter a state of moral dumbfounding (see above) in response to scenarios featuring taboo cases (such as sibling incest) even when most plausible reasons against the taboo action have been ruled out by the details of the vignette (such as consensual sex between siblings with birth control and in secret). However, rather than characterizing our aversive responses to such taboo cases as misguided, they could be considered well attuned to the fact that the two siblings played Russian roulette with their emotional health—the fact that everything turned out fine does not justify judging their risky action as "okay" (cf. Jacobson, 2012; for criticisms of the moral dumbfounding effect, see Guglielmo, 2018; Royzman, Kim, & Leeman, 2015). In a similar vein, Allman and Woodward (2008; see also Woodward & Allman, 2007) argue that our affective intuitions often provide us with important information in moral settings, such as signaling the intentions of others and their likely responses to our actions, and they can thereby lead to better moral decision making. Their view is underscored by findings demonstrating that people with damages in areas crucial to emotional processing exhibit moral behavior and decision making that would not be considered appropriate by any reasonable standard (e.g., Damasio, 1994).

The view that moral reasoning can "migrate" into people's moral intuitions is now gaining momentum (Sauer, 2017). To give a concrete example of how a rational learning process can lead to moral intuitions, let us consider a study by Nichols and colleagues (2016), who have argued that the acquisition of deontological rules may be based on a Bayesian principle called the "size principle." To illustrate this principle, imagine that there are four fair dice, each with a different denomination: 4, 6, 8, and 10. Out of your sight, one dice is picked at random and rolled 10 times, and you are told the outcomes: 2, 2, 1, 4, 3, 1, 2, 2, 3, and 4. Your task is to figure out which dice has been rolled. The formula of the size principle provides exact probabilities for the four hypotheses (4, 6, 8, 10) and favors the narrowest one (4), which corresponds to our intuitive assessment that we would expect some outcomes to be greater than 4 if a dice with a denomination greater than 4 was rolled. The size principle is able to explain children's fast learning of word categories: a few positive examples of different dogs that are called "dogs" are sufficient to infer that the term refers to dogs and not to a broader category of animals, and some examples of Dalmatians are sufficient to let them infer that the term "Dalmatian" refers to this breed of dogs and not to dogs in general (again favoring the narrower hypothesis). Nichols and colleagues provide experimental evidence indicating that deontological principles, such as the act–allow and the intended–foreseen distinction, can be learned by the same mechanism. Moreover, given the learning input children receive (as indicated by a corpus analysis), they would naturally acquire such deontological rules by approximating Bayesian inference—it might even be statistically irrational for them to instead infer utilitarian rules. Hence, Nichols and colleagues provide a rational learning mechanism for the acquisition of deontological principles that can undercut attempts to characterize deontological rules and principles as the product of irrationality. Moreover, it also offers an empiricist explanation of how rational learners can retrieve sophisticated moral rules from seemingly limited information in a way that escapes nativist arguments from the poverty of moral stimuli.

Let us assume, for the sake of the argument, that intuitions are the output of rational learning mechanisms. Does this fact mean that we should stop worrying that moral intuitions might be misleading and grant them moral authority? Some authors believe that such an optimistic view would not be justified. For instance, Gjesdal (2018) points out that since the learning mechanisms underlying our moral intuitions are domain general, they are potentially sensitive to morally relevant features—but they might also pick up morally irrelevant or even immoral information.[4] We should not consider them as morally authoritative on their own, Gjesdal argues, as long as we cannot rule out the possibility that moral intuitions are affected by morally nonrelevant features. This skeptical stance gains traction by the plausible assumption that our learning systems are well attuned and sensitive to information that is important for an agent's welfare. The reason is the—unfortunate—fact that what promotes an agent's welfare and what morality requires from her often come into conflict.[5] Hence, it cannot be ruled out that our moral intuitions are tainted by self-interested considerations. One might

address this worry by pointing out that in an appropriately structured environment, our learning mechanisms and moral intuitions could become properly attuned to morally relevant features. For instance, we might learn that a certain degree of impartiality is necessary for us to successfully navigate through the social world, and prejudices might be reduced by coming into contact with the respective people. But making the case for moral intuitions in this way would let it rest on shaky ground, since it is far from obvious that our environment was, is, or will become structured in a way that would guarantee that our moral intuitions are (or will become) adapted in a morally appropriate way. Hence, Gjesdal concludes, the finding that our intuitions are rationally attuned to reality does not warrant that they are morally appropriate. Greene (2017) seems to agree with this assessment and also points out that no matter how sophisticated learning mechanisms might be, they crucially depend on the input they get and the feedback they receive—as the saying goes in contexts that deal with data: garbage in, garbage out. He thus maintains his rather pessimistic view and concludes that moral intuitions might offer reasonable advice in our everyday social life, such as coordinating social interactions within a group ("Me versus Us," as Greene puts it), but are likely to fail when it comes to solving problems between groups with competing interests ("Us versus Them"), the moral issues surrounding novel or unfamiliar problems that originate in distinctively modern institutions, practices or technologies, or making progress in moral philosophy.

What can we conclude from these findings and reflections? On the one hand, the worry that people's moral judgments are not rational *because* they are based on intuitions seems misguided. Intuitions can be the output of highly sophisticated learning systems, and they probably have the potential to reflect moral truths and to successfully navigate us through, and solve problems in, the social world. On the other hand, there is no moral safeguard in these systems ensuring that only morally relevant information is picked up and that only morally adequate outputs are produced. Hence, the prospects of the rational-learning strategy for vindicating people's moral intuitions seem to depend on the extent to which the sophisticated internal learning mechanisms shaping our moral intuitions can avail themselves of morally relevant information in the external social environment (Gjesdal, 2018; Kumar, 2016). Investigating the conditions under which this endeavor can succeed will probably (and hopefully) be a major focus of future moral psychology.

Notes

1. To avoid misunderstandings, we do not mean to imply that the studies presented in the normative part were designed with the purpose of answering normative questions (although some may have been). Rather, most of them can be characterized as descriptive work that has more less direct implications for normative issues.

2. Unfortunately, the labels "consequentialist" and "deontological" are often used in a crude and loose way in moral psychology (e.g., intervening in Switch is actually consistent with most deontological moral theories). See Sinnott-Armstrong (2019) and Alexander and Moore (2016) for a comprehensive review of consequentialist and deontological ethics, respectively.

3. Crockett (2013) explains such means–versus–side effect cases by postulating an interaction of the model-based system and a third, Pavlovian system, which responds reflexively to aversive and rewarding states.

4. Railton (2017) reviews research specifically on moral learning. See also a recent special issue of *Cognition* on moral learning (Cushman, Kumar, & Railton, 2017).

5. Fehige and Wessels (see chapter 12.1 in this handbook) discuss at greater length the relationship between morality and an instrumental view of rationality and conclude that they cannot be brought into full harmony.

References

Alexander, L., & Moore, M. (2016). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/

Allman, J., & Woodward, J. (2008). What are moral intuitions and why should we care about them? A neurobiological perspective. *Philosophical Issues*, *18*, 164–185.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*, 74–85.

Bengson, J., Cuneo, T., & Shafer-Landau, R. (2019). Trusting moral intuitions. *Noûs*, *54*(4).

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329.

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

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241–265.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science*, *25*(2), 85–90.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In T. K. Shackelford & R. D. Hansen (Eds.),*The evolution of morality* (pp. 27–51). Cham, Switzerland: Springer.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7.

Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Psychological and philosophical perspectives. *Cognition*, *167*, 1–10.

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

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

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Enoch, D. (2011). *Taking morality seriously: A defense of robust realism*. Oxford, England: Oxford University Press.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Gerrans, P., & Kennett, J. (2010). Neurosentimentalism and moral agency. *Mind*, *119*(475), 585–614.

Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 1–26). Cambridge, MA: MIT Press.

Gilligan, C. (1982). *In a different voice*. Cambridge, MA: Harvard University Press.

Gjesdal, A. (2018). Moral learning, rationality, and the unreliability of affect. *Australasian Journal of Philosophy*, *96*, 460–473.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Graham, J., Haidt, J., & Nosek, B. A. (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. 3. The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–79). Cambridge, MA: MIT Press.

Greene, J. D. (2010). Notes on "The normative insignificance of neuroscience" by Selim Berker. Retrieved from https://static1.squarespace.com/static/54763f79e4b0c4e55ffb000c/t/54cb945ae4b001aedee69e81/1422627930781/notes-on-berker.pdf

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY: Penguin.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*(4), 695–726.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77.

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., Darley, J. M., & Cohen, J. D. (2004). The neural basis 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. (2018). Unfounded dumbfounding: How harm and purity undermine evidence for moral dumbfounding. *Cognition*, *170*, 334–337.

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

Haidt, J., & Björklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 181–217). Cambridge, MA: MIT Press.

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21.

Hauser, M. D., Young, L., & Cushman, F. (2008). Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 107–144). Cambridge, MA: MIT Press.

Huemer, M. (2008). Revisionary intuitionism. *Social Philosophy and Policy*, *25*(1), 368–392.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS One*, *7*(8), e42366.

Jacobson, D. (2012). Moral dumbfounding and moral stupefaction. *Oxford Studies in Normative Ethics*, *2*, 289–316.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm:

A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2011). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognition and Affective Neuroscience*, *7*(4), 393–402.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus & Giroux.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, *10*(2), 95–118.

Kennett, J., & Fine, C. (2009). Will the real moral judgment please stand up? *Ethical Theory and Moral Practice*, *12*(1), 77–96.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–329.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 441–447). Cambridge, MA: MIT Press.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago, IL: Rand McNally.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109.

Kumar, V. (2016). The empirical identity of moral judgment. *Philosophical Quarterly*, *66*(265), 783–804.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518–536.

Machery, E. (2014). In defense of reverse inference. *British Journal for the Philosophy of Science*, *65*(2), 251–267.

May, J. (2014a). Does disgust influence moral judgment? *Australasian Journal of Philosophy*, *92*(1), 125–141.

May, J. (2014b). Moral judgment and deontology: Empirical developments. *Philosophy Compass*, *9*(11), 745–755.

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.

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*. Cambridge, England: Cambridge University Press.

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707–718.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*(3), 573–587.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125.

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

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H. Y. (2016). Rational learners and moral rules. *Mind & Language*, *31*, 530–554.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*, 148–158.

Piaget, J. (1932). *The moral judgment of the child*. Oxford, England: Harcourt, Brace.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*(1), 29–43.

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

Prinz, J. (2008). Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 157–170). Cambridge, MA: MIT Press.

Prinz, J. (2016). Sentimentalism and the moral brain. In S. M. Liao (Ed.), *Moral brains: The neuroscience of morality* (pp. 45–74). Oxford, England: Oxford University Press.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, *124*, 813–859.

Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition*, *167*, 172–190.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment & Decision Making*, *10*(4), 296–313.

Sauer, H. (2012a). Educated intuitions: Automaticity and rationality in moral judgment. *Philosophical Explorations*, *15*(3), 255–275.

Sauer, H. (2012b). Morally irrelevant factors: What's left of the dual-process model of moral cognition? *Philosophical Psychology*, *25*(6), 783–811.

Sauer, H. (2012c). Psychopaths and filthy desks: Are emotions necessary and sufficient for moral judgment? *Ethical Theory and Moral Practice*, *15*(1), 95–115.

Sauer, H. (2015). Can't we all disagree more constructively? Moral foundations, moral reasoning, and political disagreement. *Neuroethics*, *8*(2), 153–169.

Sauer, H. (2017). *Moral judgments as educated intuitions*. Cambridge, MA: MIT Press.

Sauer, H., & Bates, T. (2013). Chairmen, cocaine, and car crashes: The Knobe effect as an attribution error. *Journal of Ethics*, *17*(4), 305–330.

Sinnott-Armstrong, W. (2019). Consequentialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2019/entries/consequentialism/

Smetana, J. G., & Braeges, J. L. (1990). The development of toddlers' moral and conventional judgments. *Merrill Palmer Quarterly*, *36*(3), 329–346.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–541.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*, 126–134.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

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

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York, NY: Oxford University Press.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*, 28–43.

Woodward, J., & Allman, J. (2007). Moral intuition: Its neural substrates and normative significance. *Journal of Physiology*, *101*(4), 179–202.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786–791.

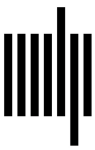
The Handbook of Rationality

Edited by: Markus Knauff, Wolfgang Spohn

Citation:

The Handbook of Rationality
Edited by: Markus Knauff, Wolfgang Spohn
DOI: 10.7551/mitpress/11252.001.0001
ISBN (electronic): 9780262366175
Publisher: The MIT Press
Published: 2021

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Knauff, Markus, editor. | Spohn, Wolfgang, editor.
Title: The handbook of rationality / edited by Markus Knauff and Wolfgang Spohn.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020048455 | ISBN 9780262045070 (hardcover)
Subjects: LCSH: Reasoning (Psychology) | Reason. | Cognitive psychology. | Logic. | Philosophy of mind.
Classification: LCC BF442 .H36 2021 | DDC 153.4/3—dc23
LC record available at https://lccn.loc.gov/2020048455