Morality and Cognitive Science

What do we know about how people make moral judgments? And what should moral philosophers do with this knowledge? This article addresses the cognitive science of moral judgment. It reviews important empirical findings and discusses how philosophers have reacted to them.

Several trends have dominated the cognitive science of morality in the early 21st century. One is a move away from strict opposition between biological and cultural explanations of morality's origin, toward a hybrid account in which culture greatly modifies an underlying common biological core. Another is the fading of strictly rationalist accounts in favor of those that recognize an important role for unconscious or heuristic judgments. Along with this has come expanded interest in the psychology of reasoning errors within the moral domains. Another trend is the recognition that moral judgment interacts in complex ways with judgment in other domains; rather than being caused by judgments about intention or free will, moral judgment may partly influence them. Finally, new technology and neuroscientific techniques have led to novel discoveries about the functional organization of the moral brain and the roles that neurotransmitters play in moral judgment.

Philosophers have responded to these developments in a variety of ways. Some deny that the cognitive science of moral judgment has any relevance to philosophical reflection on how we ought to live our lives, or on what is morally right to do. One argument to this end follows the traditional is/ought distinction and insists that we cannot generate a moral ought from any psychological is. Another argument insists that the study of morality is autonomous from scientific inquiry, because moral deliberation is essentially first-personal and not subject to any third-personal empirical correction.

Other philosophers argue that the cognitive science of moral judgment may have significant revisionary consequences for our best moral theories. Some make an epistemic argument: if moral judgment aims at discovering moral truth, then psychological findings can expose when our judgments are unreliable, like faulty scientific instruments. Other philosophers focus on non-epistemic factors, such as the need for internal consistency within moral judgment, the importance of conscious reflection, or the centrality of intersubjective justification. Certain cognitive scientific findings might require a new approach to these features of morality.
The first half of this article (sections 1 to 4) surveys the cognitive science literature, describing key experimental findings and psychological theories in the moral domain. The second half (sections 5 to 10) discusses how philosophers have reacted to these findings, discussing different ways philosophers have sought to immunize moral inquiry from empirical revision, or enthusiastically taken up psychological tools to make new moral arguments.

Note that the focus of this article is on moral judgment. See the article “Moral Character” for discussion of the relationship between cognitive science and moral character.

Table of Contents

1. Biological and Cultural Origins of Moral Judgment
2. The Psychology of Moral Reasoning
3. Interaction between Moral Judgment and Other Cognitive Domains
4. The Neuroanatomy of Moral Judgment
5. What Do Moral Philosophers Think of Cognitive Science?
6. Moral Cognition and the Is/Ought Distinction
   a. Semantic Is/Ought
   b. Non-semantic Is/Ought
7. The Autonomy of Morality
8. Moral Cognition and Moral Epistemology
9. Non-epistemic Approaches
   a. Consistency in Moral Reasoning
   b. Rational Agency
   c. Intersubjective Justification
10. Objections and Alternatives
    a. Explanation and Justification
    b. The Expertise Defense
    c. The Regress Challenge
    d. Positive Alternatives
11. References and Further Reading

1. Biological and Cultural Origins of Moral Judgment

One key empirical question is this: are moral judgments rooted in innate factors or are they fully explained by acquired cultural traits? During the 20th century, scientists tended to adopt extreme positions on this question. The psychologist B. F. Skinner (1971) saw moral rules as socially conditioned patterns of behavior; given the right reinforcement, people could be led to judge virtually anything morally right or wrong. The biologist E. O. Wilson (1975), in contrast, argued that nearly all of human morality could be understand via the application of evolutionary biology.
Around the early 21st century, however, most researchers on moral judgment began to favor a hybrid model, allowing roles for both biological and cultural factors.

There is evidence that at least the precursors of moral judgment are present in humans at birth, suggesting an evolutionary component. In a widely cited study, Kiley Hamlin and colleagues examined the social preferences of pre-verbal infants (Hamlin, Wynn, and Bloom 2007). The babies, seated on their parents’ laps, watched a puppet show in which a character trying to climb a hill was helped up by one puppet, but pushed back down by another puppet. Offered the opportunity to reach out and grasp one of the two puppets, babies showed a preference for the helping puppet over the hindering puppet. This sort of preference is not yet full-fledged moral judgment, but it is perhaps the best we can do in assessing the social responses of humans before the onset of language, and it suggests that however human morality comes about, it builds upon innate preferences for pro-social behavior.

A further piece of evidence comes from the work of theorists Leda Cosmides and John Tooby (1992), who argue that the minds of human adults display an evolutionary specialization for moral judgment. The Wason Selection Task is an extremely well-established paradigm in the psychology of reasoning, which shows that most people make persistent and predictable mistakes in evaluating abstract inferences. A series of studies by Cosmides, Tooby, and their colleagues shows that people do much better on a form of this task when it is restricted to violations of social norms. So, for instance, rather than being asked to evaluate an abstract rule linking numbers and colors, people were asked to evaluate a rule prohibiting those below a certain age from consuming alcohol. Participants in these studies made the normal mistakes when looking for violations of abstract rules, but made fewer mistakes in detecting violations of social rules. According to Cosmides and Tooby, these results suggest that moral judgment evolved as a domain-specific capacity, rather than an as application of domain-general reasoning. If this is right, then there must be at least an innate core to moral judgments.

Perhaps the most influential hybrid model of innate and cultural factors in moral judgment research is the linguistic analogy research program (Dwyer 2006; Hauser 2006; Mikhail 2011). This approach is explicitly modeled on Noam Chomsky’s (1965) generative linguistics. According to Chomsky, the capacity for language production and some basic structural parameters for functioning grammar are innate, but the enormous diversity of human languages comes about through myriad cultural settings and prunings within the evolutionarily allowed range of possible grammars. By analogy, then, moral grammar suggests that the capacity of making moral judgments is innate, along with some basic constraints on the substance of the moral domain—but within this evolutionarily enabled space, culture works to select and prune distinct local moralities.

The psychologist Jonathan Haidt (2012) has highlighted the role of cultural difference in the scientific study of morality through his influential Moral Foundations account. According to Haidt,
all documented moral beliefs can be classified as fitting within a handful of moral sub-domains, such as harm-aversion, justice (in distribution of resources), respect for authority, and purity. Haidt argues that moral differences between cultures reflect differences in emphasis on these foundations. In fact, he suggests, industrialized western cultures appear to have emphasized the harm-aversion and justice foundations almost exclusively, unlike many other world cultures. Yet Haidt also insists upon a role for biology in explaining moral judgment; he sees each of the foundations as rooted in a particular evolutionary origin. Haidt’s foundations account remains quite controversial, but it is a prominent example of contemporary scientists’ focus on avoiding polarized answers to the biology versus culture question.

The rest of this article mostly sets aside further discussion of the distal—evolutionary or cultural—explanation of moral judgment. Instead it focuses on proximal cognitive explanations for particular moral judgments. This is because the ultimate aim is to consider the philosophical significance of moral cognitive science, whereas the moral philosophical uptake of debates over evolution is discussed elsewhere. Interested readers can see the articles on “Evolutionary Ethics” and “Moral Relativism.”

2. The Psychology of Moral Reasoning

One crucial question is whether moral judgments arise from conscious reasoning and reflection, or are triggered by unconscious and immediate impulses. In the 1970s and 1980s, research in moral psychology was dominated by the work of Lawrence Kohlberg (1971), who advocated a strongly rationalist conception of moral judgment. According to Kohlberg, mature moral judgment demonstrates a concern with reasoning through highly abstract social rules. Kohlberg asked his participants (boys and young men) to express opinions about ambiguous moral dilemmas and then explain the reasons behind their conclusions. Kohlberg took it for granted that his participants were engaged in some form of reasoning; what he wanted to find out was the nature and quality of this reasoning, which he claimed emerged through a series of developmental stages. This approach came under increasing criticism in the 1980s, particularly through the work of feminist scholar Carol Gilligan (1982), who exposed the trouble caused by Kohlberg’s reliance on exclusively male research participants. See the article “Moral Development” for further discussion of that issue.

Since the turn of the twentieth century psychologists have placed much less emphasis on the role of reasoning in moral judgment. A turning point was the publication of Jonathan Haidt’s paper, “The Emotional Dog and Its Rational Tail” (Haidt 2001). There Haidt discusses a phenomenon he calls “moral dumbfounding.” He presented his participants with provocative stories, such as a brother and sister who engage in deliberate incest, or a man who eats his (accidentally) dead dog. Asked to explain their moral condemnation of these characters, participants cited reasons that seem to be ruled out by the description of the stories—for instance, participants said that the incestuous siblings might create a child with dangerous birth defects, though the original story made clear that
they took extraordinary precautions to avoid conception. When reminded of these details, participants did not withdraw their moral judgments; instead they said things like, “I don’t know why it’s wrong, it just is.” According to Haidt, studies of moral dumbfounding confirm a pattern of evidence that people do not really reason their way to moral conclusions. Moral judgment, Haidt argues, arrives in spontaneous and unreflective flashes, and reasoning is only something done post hoc, to rationalize the already-made judgments.

Many scientists and philosophers have written about the evidence Haidt discusses. A few take extremist positions, absolutely upholding the old rationalist tradition or firmly insisting that reasoning plays no role at all. (Haidt himself has slightly softened his anti-rationalism in later publications.) But it is probably right to say that the dominant view in moral psychology is a hybrid model. Many of our everyday moral judgments do arise in sudden flashes, without any need for explicit reasoning. But when faced with new and difficult moral situations, and sometimes even when confronted with explicit arguments against our existing beliefs, we are able to reason our way toward new moral judgments.

The dispute between rationalists and their antagonists is primarily one about procedure: are moral judgments produced through conscious thought or unconscious psychological mechanisms? There is a related but distinct set of issues concerning the quality of moral judgment. However moral judgment works, explicitly or implicitly, is it rational? Is the procedure that produces our moral judgments a reliable procedure? (It is important to note that a mental procedure might be unconscious and still be rational. Many of our simple mathematical calculations are accomplished unconsciously, but that alone does not keep them from counting as rational.)

There is considerable evidence suggesting that our moral judgments are unreliable (and so arguably irrational). Some of this evidence comes from experiments imported from other domains of psychology, especially from behavioral economics. Perhaps most famous is the Asian disease case first employed by the psychologists Amos Tversky and Daniel Kahneman in the early 1980s. Here is the text they gave to some of their participants:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat this disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

If Program A is adopted, 200 people will be saved.

If Program B is adopted, there is a 1/3 probability that 600 will be saved, and 2/3 probability that no people will be saved. (Tversky and Kahneman 1981, 453)
Other participants read a modified form of this scenario, with the first option being that 400 people will die and the second option being that there is a $1/3$ probability that nobody will die, and a $2/3$ probability that 600 people will die. Notice that this is a change only in wording: the first option is the same either way, whether you describe it as 200 people being saved or 400 dying, and the second option gives the same probabilities of outcomes in either description. Notice also that, in terms of probability-expected outcome, the two programs are mathematically equivalent (for expected values, certain survival of $1/3$ of people is equivalent to a $1/3$ chance of all people surviving). Yet participants responded strongly to the way the choices are described; those who read the original phrasing preferred Program A three to one, while those who read the other phrasing showed almost precisely the opposite preference. Apparently, describing the choice in terms of saving makes people prefer the certain outcome (Program A) while describing it in terms of dying makes people prefer the chancy outcome (Program B).

This study is one of the most famous in the literature on framing effects, which shows that people’s judgments are affected by merely verbal differences in how a set of options is presented. Framing effects have been shown in many types of decision, especially in economics, but when the outcomes concern the deaths of large numbers of people it is clear that they are of moral significance. Many studies have shown that people’s moral judgments can be manipulated by mere changes in verbal description, without (apparently) any difference in features that matter to morality (see Sinnott-Armstrong 2008 for other examples).

Partly on this basis, some theorists have advocated a heuristics and biases approach to moral psychology. A cognitive bias is a systematic defect in how we think about a particular domain, where some feature influences our thinking in a way that it should not. Some framing effects apparently trigger cognitive biases; the saving/dying frame exposes a bias toward risk-taking in save frames and a bias away from it in dying frames. (Note that this leaves open whether either is the correct response—the point is that the mere difference of frame shouldn’t affect our responses, so at least one of the divergent responses must be mistaken.)

In the psychology of (non-moral) reasoning, a heuristic is a sort of mental short-cut, a way of skipping lengthy or computationally demanding explicit reasoning. For instance, if you want to know which of several similar objects is the heaviest, you could assume that it is the largest. Heuristics are valuable because they save time and are usually correct—but in certain types of cases an otherwise valuable heuristic will make predictable errors (some small objects are actually denser and so heavier than large objects). Perhaps some of our simple moral rules (“do no harm”) are heuristics of this sort—right most of the time, but unreliable in certain cases (perhaps it is sometimes necessary to harm people in emergencies). Some theorists (for example, Sunstein 2005) argue that large sectors of our ordinary moral judgments can be shown to exhibit heuristics and biases. If this is right, most moral judgment is systematically unreliable.
A closely related type of research shows that moral judgment is affected not only by irrelevant features within the questions (such as phrasing) but also by completely accidental features of the environment in which we make judgments. To take a very simple example: if you sit down to record your judgments about morally dubious activities, the severity of your reaction will depend partly on the cleanliness of the table (Schnall et al. 2008). You are likely to give a more harsh assessment of the bad behavior if the table around you is sticky and covered in pizza boxes than if it is nice and clean. Similarly, you are likely to render more negative moral judgments if you have been drinking bitter liquids (Eskine, Kacinik, and Prinz 2011), or handling dirty money (Yang et al. 2013). Watching a funny movie will make you temporarily less condemning of violence (Strohminger, Lewis, and Meyer 2011).

Some factors affecting moral judgment are internal to the judge rather than the environment. If you’ve been given a post-hypnotic suggestion to feel queasy whenever you hear a certain completely innocuous word, you will probably make more negative moral judgments about characters in stories containing those triggering words (Wheatley and Haidt 2005). Such effects are not restricted to laboratories; one Israeli study showed that parole boards are more likely to be lenient immediately after meals than when hungry (Danziger, Levav, and Avnaim-Pesso 2011).

Taken together, these studies appear to show that moral judgment is affected by factors that are morally irrelevant. The degree of wrongness of an act does not depend on which words are used to describe it, or the cleanliness of the desk at which it is considered, or whether the thinker has eaten recently. There seems to be strong evidence, then, that moral judgments are at least somewhat unreliable. Moral judgment is not always of high quality. The extent of this problem, and what difference it might make to philosophical morality, is a matter that is discussed later in the article.

3. Interaction between Moral Judgment and Other Cognitive Domains

Setting aside for now the reliability of moral judgment, there are other questions we can ask about its production, especially about its causal structure. Psychologically speaking, what factors go into producing a particular moral judgment about a particular case? Are these the same factors that moral philosophers invoke when formally analyzing ethical decisions? Empirical research appears to suggest otherwise.

One example of this phenomenon concerns intention. Most philosophers have assumed that in order for something done by a person to be evaluable as morally right or wrong, the person must have acted intentionally (at least in ordinary cases, setting aside negligence). If you trip me on purpose, that is wrong, but if you trip me by accident, that is merely unfortunate. Following this intuitive point, we might think that assessment of intentionality is causally prior to assessment of morality. That is, when I go to evaluate a potentially morally important situation, I first work out
whether the person involved acted intentionally, and then use this judgment as an input to working out whether what they have done is morally wrong. Hence, in this simple model, the causal structure of moral judgment places it downstream from intentionality judgment.

But empirical evidence suggests that this simple model is mistaken. A very well-known set of studies concerning the side-effect effect (also known as the Knobe effect, for its discoverer Joshua Knobe) appears to show that the causal relationship between moral judgment and intention-assessment is much more complicated. Participants were asked to read short stories like the following:

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. (Knobe 2003, 191)

Other participants read the same story, except that the program would instead help, rather than harm, the environment as a side effect. Both groups of participants were asked whether the executive intentionally brought about the side effect. Strikingly, when the side effect was a morally bad one (harming the environment), 82% of participants thought it was brought about intentionally, but when the side effect was morally good (helping the environment) 77% of participants thought it was not brought about intentionally.

There is a large literature offering many different accounts of the side-effect effect (which has been repeatedly experimentally replicated). One plausible account is this: people sometimes make moral judgments prior to assessing intentionality. Rather than intentionality-assessment always being an input to moral judgment, sometimes moral judgment feeds input to assessing intentionality. A side effect judged wrong is more likely to be judged intentional than one judged morally right. If this is right, then the simple model of the causal structure of moral judgment cannot be quite correct—the causal relation between intentionality-assessment and moral judgment is not unidirectional.

Other studies have shown similar complexity in how moral judgment relates to other philosophical concepts. Philosophers have often thought that questions about causal determinism and free will are conceptually prior to attributing moral responsibility. That is, whether or not we can hold people morally responsible depends in some way on what we say about freedom of the will. Some philosophers hold that determinism is compatible with moral responsibility and others deny this, but both groups start from thinking about the metaphysics of agency and move toward morality. Yet experimental research suggests that the causal structure of moral judgment works somewhat differently (Nichols and Knobe 2007). People asked to judge whether an agent can be morally responsible in a deterministic universe seem to base their decision in part on how strongly they
Physiological approaches to the study of moral judgment have taken on an increasingly important role. Employing neuroscientific and psychopharmacological research techniques, these studies help to illuminate the functional organization of moral judgment by revealing the brain areas implicated in its exercise.

An especially central concern in this literature is the role of emotion in moral judgment. An early influential study by Jorge Moll and colleagues (2005) demonstrated selective activity for moral judgment in a network of brain areas generally understood to be central to emotional processing. This work employed functional magnetic resonance imaging (fMRI), the workhorse of modern neuroscience, in which a powerful magnet is used to produce a visual representation of relative levels of cellular energy used in various brain areas. Employing fMRI allows researchers to get a near-real-time representation of the brain’s activities while the conscious subject makes judgments about morally important scenarios.

One extremely influential fMRI study of moral judgment was conducted by philosopher-neuroscientist Joshua D. Greene and colleagues. They compared brain activity in people making deontological moral judgments with brain activity while making utilitarian moral judgments. (To oversimplify: a utilitarian moral judgment is one primarily attentive to the consequences of a morally evaluate the agent’s actions. Scenarios involving especially vivid and egregious moral violations tend to produce greater endorsement of compatibilism than more abstract versions. The interpretation of these studies is highly controversial, but at minimum they seem to cast doubt on a simple causal model placing judgments about free will prior to moral judgment.

A similar pattern holds for judgments about the true self. Some moral philosophers hold that morally responsible action is action resulting from desires or commitments that are a part of one’s true or deep self, rather than momentary impulses or external influences. If this view reflects how moral judgments are made, then we should expect people to first assess whether a given action results from an agent’s true self and then render moral judgment about those that do. But it turns out that moral judgment provides input to true self assessments. Participants in one experiment (Newman, Bloom, and Knobe 2014) were asked to consider a preacher who explicitly denounced homosexuality while secretly engaging in gay relationships. Asked to determine which of these behaviors reflected the preacher’s true self, participants first employed their own moral judgment; those disposed to see homosexuality as morally wrong thought the preacher’s words demonstrated his true self, while those disposed to accept homosexuality thought the preacher’s relationships came from his true self. The implication is that moral judgment is sometimes a causal antecedent of other types of judgments, including those that philosophers have thought conceptually prior to assessing morality.

4. The Neuroanatomy of Moral Judgment

Physiological approaches to the study of moral judgment have taken on an increasingly important role. Employing neuroscientific and psychopharmacological research techniques, these studies help to illuminate the functional organization of moral judgment by revealing the brain areas implicated in its exercise.

An especially central concern in this literature is the role of emotion in moral judgment. An early influential study by Jorge Moll and colleagues (2005) demonstrated selective activity for moral judgment in a network of brain areas generally understood to be central to emotional processing. This work employed functional magnetic resonance imaging (fMRI), the workhorse of modern neuroscience, in which a powerful magnet is used to produce a visual representation of relative levels of cellular energy used in various brain areas. Employing fMRI allows researchers to get a near-real-time representation of the brain’s activities while the conscious subject makes judgments about morally important scenarios.

One extremely influential fMRI study of moral judgment was conducted by philosopher-neuroscientist Joshua D. Greene and colleagues. They compared brain activity in people making deontological moral judgments with brain activity while making utilitarian moral judgments. (To oversimplify: a utilitarian moral judgment is one primarily attentive to the consequences of a
decision, even allowing the deliberate sacrifice of an innocent to save a larger number of others. Deontological moral judgment is harder to define, but for our purposes means moral judgment that responds to factors other than consequences, such as the individual rights of someone who might be sacrificed to save a greater number. See “Ethics.”

In a series of empirical studies and philosophical papers, Greene has argued that his results show that utilitarian moral judgment correlates with activity in cognitive or rational brain areas, while deontological moral judgment correlates with activity in emotional areas (Greene 2008). (He has since softened this view a bit, conceding that both types of moral judgment allow some form of emotional processing. He now holds that deontological emotions are a type that trigger automatic behavioral responses, whereas utilitarian emotions are flexible prompts to deliberation (Greene 2014).) According to Greene, learning these psychological facts give us reason to distrust our deontological judgments; in effect, his is a neuroscience-fueled argument on behalf of utilitarianism. This argument is at the center of a still very spirited debate. Whatever its outcome, Greene’s research program has had an undeniable influence on moral psychology; his scenarios (which derive from philosophical thought experiments by Philippa Foot and Judith Jarvis Thomson) have been adopted as standard across much of the discipline, and the growth of moral neuroimaging owes much to his project.

Alongside neuroimaging, lesion study is one of the central techniques of neuroscience. Recruiting research participants who have pre-existing localized brain damage (often due to physical trauma or stroke) allows scientists to infer the function of a brain area from the behavioral consequences of its damage. For example, participants with damage to the ventromedial prefrontal cortex, who have persistent difficulties with social and emotional processing, were tested on dilemmas similar to those used by Greene (Koenigs et al. 2007). These patients show a greater tendency toward utilitarian judgments than did healthy controls. Similar lesion studies have since found a range of different results, so the empirical debate remains unsettled, but the technique continues to be important.

Two newer neuroscientific research techniques have begun to play important roles in moral psychology. Transcranial magnetic simulation (TMS) uses blasts of electromagnetism to suppress or heighten activity in a brain region. In effect, this allows researchers to (temporarily) alter healthy brains and correlate this alteration with behavioral effects. For instance, one study (Young et al. 2010) used TMS to suppress activity in the right temporoparietal junction, an area associated with assessing the mental states of others (see “Theory of Mind.” After the TMS treatment, participants’ moral judgments showed less sensitivity to whether characters in a dilemma acted intentionally or accidentally. Another technique, transcranial direct current stimulation (TCDS) has been shown to increase compliance with social norms when applied to the right lateral prefrontal cortex (Ruff, Ugazio, and Fehr 2013).
Finally, it is possible to study the brain not only at the gross structural scale, but also by examining its chemical operations. Psychopharmacology is the study of the cognitive and behavioral effects of chemical alteration of the brain. In particular, the levels of neurotransmitters, which regulate brain activity in a number of ways, can be manipulated by introducing pharmaceuticals. For example, participants’ readiness to make utilitarian moral judgments can be altered by administration of the drugs propranolol (Terbeck et al. 2013) and citalopram (Crockett et al. 2010).

5. What Do Moral Philosophers Think of Cognitive Science?

So far this article has described the existing science of moral judgment: what we have learned empirically about the causal processes by which moral judgments are produced. The rest of the article discusses the philosophical application of this science. Moral philosophers try to answer substantive ethical questions: what are the most valuable goals we could pursue? How should we resolve conflicts among these goals? Are there ways we should not act even if doing so would promote the best outcome? What is the shape of a good human life and how could we acquire it? How is a just society organized? And so on.

What cognitive science provides is causal information about how we typically respond to these kinds of questions. But philosophers disagree about what to make of this information. Should we ever change our answers to ethical questions on the basis of what we learn about their causal origins? Is it ever reasonable (or even rationally mandatory) to abandon a confident moral belief because of a newly learned fact about how one came to believe it?

We now consider several prominent responses to these questions. We start with views that deny much or any role for cognitive science in moral philosophy. We then look at views that assign to cognitive science a primarily negative role, in disqualifying or diminishing the plausibility of certain moral beliefs. We conclude by examining views that see the findings of cognitive science as playing a positive role in shaping moral theory.

6. Moral Cognition and the Is/Ought Distinction

We must start with the famous is/ought distinction. Often attributed to Hume (see “Hume”), the distinction is a commonsensical point about the difference between descriptive claims that characterize how things are (for example, “the puppy was sleeping”) and prescriptive claims that assert how things should or should not be (for instance, “it was very wrong of you to kick that sleeping puppy”). These are widely taken to be two different types of claims, and there is a lot to be said about the relationship between them. For our purposes, we may gloss it as follows: people often make mistakes when they act as if an ought-claim follows immediately and without further argument from an is-claim. The point is not (necessarily) that it is always a mistake to draw an
ought-statement as the conclusion of an is-statement, just that the relationship between them is messy and it is easy to get confused. Some philosophers do assert the much stronger claim that ought-statements can never be validly drawn from is-statements, but this is not what everyone means when the issue is raised.

For our purposes, we are interested in how the is/ought distinction might matter to applying cognitive science to moral philosophy. Cognitive scientific findings are is-type claims; they describe facts about how our minds actually do work, not about how they ought to work. Yet the kind of cognitive scientific claims at interest here are claims about moral cognition—is-claims about the origin of ought-claims. Not surprisingly then, if the is/ought distinction tends to mark moments of high risk for confusion, we should expect this to be one of those moments. Some philosophers have argued that attempts to change moral beliefs on the basis of cognitive scientific findings are indeed confusions of this sort.

**a. Semantic Is/Ought**

In the mid-twentieth century it was popular to understand the is/ought distinction as a point about moral semantics. That is, the distinction pointed to a difference in the implicit logic of two uses of language. Descriptive statements (is-statements) are, logically speaking, used to attribute properties to objects; “the puppy is sleeping” just means that the sleeping-property attaches to the puppy-object. But prescriptive statements (ought-statements) do not have this logical structure. Their surface descriptive grammar disguises an imperative or expressive logic. So “it was very wrong of you to kick that sleeping puppy” is not really attributing the property of wrongness to your action of puppy-kicking. Rather, the statement means something like “don’t kick sleeping puppies!” or even “kicking sleeping puppies? Boo!” Or perhaps “do not like it when sleeping puppies are kicked and I want you to not like it as well.” (See “Ethical Expressivism.”)

If this analysis of the semantics of moral statements is right, then we can easily see why it is a mistake to draw ought-conclusions from is-premises. Logically speaking, simple imperatives and expressives do not follow from simple declaratives. If you agree with “the puppy was sleeping” yet refuse to accept the imperative “don’t kick sleeping puppies!” you haven’t made a logical mistake. You haven’t made a logical mistake even if you also agree with “kicking sleeping puppies causes them to suffer.” The point here isn’t about the moral substance of animal cruelty—the point is about the logical relationship between two types of language use. There is no purely logical compulsion to accept any simple imperative or expressive on the basis of any descriptive claim, because the types of language do not participate in the right sort of logical relationship.

Interestingly, this sort of argument has not played much of a role in the debate about moral cognitive science, though seemingly it could. After all, cognitive scientific claims are, logically speaking, descriptive claims, so we could challenge their logical relevance to assessing any
imperative or expressive claims. But, historically speaking, the rise of moral cognitive science came mostly after the height of this sort of semantic argument. Understanding the is/ought distinction in semantic terms like these had begun to fade from philosophical prominence by the 1970s, and especially by the 1980s when modern moral cognitive science (arguably) began. Contemporary moral expressivists are often eager to explain how we can legitimately draw inferences from is to ought despite their underlying semantic theory.

It is possible to see the simultaneous decline of simple expressivism and the rise of moral cognitive science as more than mere coincidence. Some historians of philosophy see the discipline as having pivoted from the linguistic turn of the late 19th and early 20th century to the cognitive turn (or conceptual turn) of the late 20th century. Philosophy of language, while still very important, has receded from its position at the center of every philosophical inquiry. In its place, at least in some areas of philosophy, is a growing interest in naturalism and consilience with scientific inquiry. Rather than approaching philosophy via the words we use, theorists often now approach it through the concepts in our minds—concepts which are in principle amenable to scientific study. As philosophers became more interested in the empirical study of their topics, they were more likely to encourage and collaborate with psychologists. This has certainly contributed to the growth of moral cognitive science since the 1990s.

b. Non-semantic Is/Ought

Setting aside semantic issues, we still have a puzzle about the is/ought distinction. How do we get to an ought conclusion from an is premise? The idea that prescriptive and descriptive claims are different types of claim retains its intuitive plausibility. Some philosophers have argued that scientific findings cannot lead us to (rationally) change our moral beliefs because science issues only descriptive claims. It is something like a category mistake to allow our beliefs in a prescriptive domain to depend crucially upon claims within a descriptive domain. More precisely, it is a mistake to revise your prescriptive moral beliefs because of some purely descriptive fact, even if it is a fact about those beliefs.

Of course, the idea here cannot be that it is always wrong to update moral beliefs on the basis of new scientific information. Imagine that you are a demolition crew chief and you are just about to press the trigger to implode an old factory. Suddenly one of your crew members shouts, “Wait, look at the thermal monitor! There’s a heat signature inside the factory—that’s probably a person! You shouldn’t press the trigger!” It would be extremely unfitting for you to reply that whether or not you should press the trigger cannot depend on the findings of scientific contraptions like thermal monitors.

What this example shows is that the is/ought problem, if it is a problem, is obviously not about excluding all empirical information from moral deliberation. But it is important to note that the
scientific instrument itself does not tell us what we ought to do. We cannot just read off moral conclusions from descriptive scientific facts. We need some sort of bridging premise, something that connects the purely descriptive claim to a prescriptive claim. In the demolition crew case, it is easy to see what this bridging premise might be, something like: “if pressing a trigger will cause the violent death of an innocent person, then you should not press the trigger.” In ordinary moral interactions we often leave bridging premises implicit—it is obvious to everyone in this scenario that the presence of an innocent human implies the wrongness of going ahead with implosion.

But there is a risk in leaving our bridging premises implicit. Sometimes people seem to be relying upon implicit bridging premises that are not mutually agreed on, or that may not make any sense at all. Consider: “You shouldn’t give that man any spare change. He is a Templar.” Here you might guess that the speaker thinks Templars are bad people and do not deserve help, but you might not be sure—why would your interlocutor even care about Templars? And you are unlikely to agree with this premise anyway, so unless the person makes their anti-Templar views explicit and justifies them, you do not have much reason to follow their moral advice. Another example: “The tiles in my kitchen are purple, so it’s fine for you to let those babies drown.” It is actually hard to interpret this utterance as something other than a joke or metaphor. If someone tried to press it seriously, we would certainly demand to be informed of the bridging premise between linoleum color and nautical infanticide, and we would be skeptical that anything plausible might be provided.

So far, so simple. Now consider: “Brain area B is extremely active whenever you judge it morally wrong to cheat on your taxes. So it is morally wrong to cheat on your taxes.” What should we make of this claim? The apparent implicit bridging premise is: If brain area B gets excited when you judge it wrong to X, then it is wrong to X. But this is a very strange bridging premise. It does not make reference to any internal features of tax-cheating that might explain its wrongness. In fact, the premise appears to suggest that an act can be wrong simply because someone thinks it is wrong and some physical activity is going on inside that person’s body. This is not the sort of thing we normally offer to get to moral conclusions, and it is not immediately clear why anyone would find it convincing. Perhaps, as we said, this is an example of how easy it is to get confused about is and ought claims.

Two points come out here. First, some attempts at drawing moral conclusions from cognitive science involve implicit bridging premises that fall apart when anyone attempts to make them explicit. This is often true of popular science journalism, in which some new psychological finding is claimed to prove that such-and-such moral belief is mistaken. At times, philosophers have accused their psychologist or philosopher opponents of doing something similar. According to Berker (2009), Joshua Greene’s neuroscience-based debunking of deontology (discussed above) lacks a convincing bridging premise. Berker suggest that Greene avoids being explicit about this premise, because when it is made explicit it is either a traditional moral argument which does not
use cognitive science to motivate its conclusion or it employs cognitive scientific claims but does not lead to any plausible moral conclusion. Hence, Berker says, the neuroscience is normatively insignificant; it does not play a role in any plausible bridging premise to a moral conclusion.

Of course, even if this is right, then it implies only that Greene’s argument fails. But Berker and other philosophers have expressed doubt that any cognitive science-based argument could generate a moral conclusion. They suggest that exciting newspaper headlines and book subtitles (for example, *How Science Can Determine Human Values* (Harris 2010)) trade on our leaving their bridging premises implicit and unchallenged. For—this is the second point—if there were a successful bridging principle, it would be a very unusual one. Why, we want to know, could the fact that such-and-such causal process precedes or accompanies a moral judgment give us reason to change our view about that moral judgment? Coincident causal processes do not appear to be morally relevant. What your brain is doing while you are making moral judgments seems to be in the same category as the color of your kitchen tiles—why should it matter?

Of course, the fact that causal processes are not typically used in bridging premises does not show us that they could not. But it does perhaps give us reason to be suspicious, and to insist that the burden of proof is on anyone who wishes to present such an argument. They must explain to us why their use of cognitive science is normatively significant. A later section considers different attempts to meet this burden of proof. First, though, consider an argument that it can never be met, even in principle.

### 7. The Autonomy of Morality

Some philosophers hold that it is a mistake to try to draw moral conclusions from psychological findings because doing so misunderstands the nature of moral deliberation. According to these philosophers, moral deliberation is essentially first-personal, while cognitive science can give us only third-personal forms of information about ourselves. When you are trying to decide what to do, morally speaking, you are looking for reasons that relate your options to the values you uphold. I have moral reason not to kick puppies because I recognize value in puppies (or at least the non-suffering of puppies). Psychological claims about your brain or psychological apparatus might be interesting to someone else observing you, but they are beside the point of first-personal moral deliberation about how to act.

Here is how Ronald Dworkin makes this point. He asks us to imagine learning some new psychological fact about why we have certain beliefs concerning justice. Suppose that the evidence suggests our justice beliefs are caused by self-interested psychological processes. Then:

> It will be said that it is unreasonable for you still to think that justice requires anything, one way or the other. But why is that unreasonable? Your opinion is one about justice, not about
your own psychological processes . . . You lack a normative connection between the bleak psychology and any conclusion about justice, or any other conclusion about how you should vote or act. (Dworkin 1996, 124–125)

The idea here is not just that beliefs about morality and beliefs about psychology are about different topics. Part of the point is that morality is special—it is not just another subject of study alongside psychology and sociology and economics. Some philosophers put this point in terms of an a priori/a posteriori distinction: morality is something that we can work out entirely in our minds, without needing to go and do experiments (though of course we might need empirical information to apply moral principles once we have figured them out). Notice that when philosophers debate moral principles with one another, they do not typically conduct or make reference to empirical studies. They think hard about putative principles, come up with test cases that generate intuitive verdicts, and then think hard again about how to modify principles to make them fit to these verdicts. The process does not require empirical input, so there is no basis for empirical psychology to become involved.

This view is sometimes called the autonomy of morality (Fried 1978; Nagel 1978). It holds that, in the end, the arbiter of our moral judgments will be our moral judgments—not anything else. The only way you can get a moral conclusion from a psychological finding is to provide the normative connection that Dworkin refers to. So, for instance: if you believe that it is morally wrong to vote in a way triggered by a selfish psychological process, then finding out that your intending to vote for the Egalitarian Party is somehow selfishly motivated could give you a reason to reconsider your vote. But notice that this still crucially depends upon a moral judgment—the judgment that it is wrong to vote in a way triggered by selfish psychological processes. There is no way to get entirely out of the business of making moral judgments; psychological claims are morally inert unless accompanied by explicitly moral claims.

The point here can be made in weaker and stronger forms. The weaker form is simply this: empirical psychology cannot generate moral conclusions entirely on its own. This weaker form is accepted by most philosophers; few deny that we need at least some moral premises in order to get moral conclusions. Notice though that even the weaker form casts doubt on the idea that psychology might serve as an Archimedean arbiter of moral disagreement. Once we’ve conceded that we must rely upon moral premises to get results from psychological premises, we cannot claim that psychology is a value-neutral platform from which to settle matters of moral controversy.

The stronger form of the autonomy of morality claim holds that a psychological approach to morality fundamentally misunderstands the topic. Moral judgment, in this view, is about taking agential responsibility for our own value and actions. Re-describing these in psychological terms, so that our commitments are just various causal levers, represents an abdication of this
responsibility. We should instead maintain a focus on thinking through the moral reasons for and against our positions, leaving psychology to the psychologists.

Few philosophers completely accept the strongest form of the autonomy of morality. That is, most philosophers agree that there are at least some ways psychological genealogy could change the moral reasons we take ourselves to have. But it will be helpful for us to keep this strong form in mind as a sort of null hypothesis. As we now turn to various theoretical arguments in support of a role for cognitive science in moral theory, we can understand each as a way of addressing the strong autonomy of morality challenge. They are arguments demonstrating that psychological investigation is important to our understanding of our moral commitments.

8. Moral Cognition and Moral Epistemology

Many moral philosophers think about their moral judgments (or intuitions) as pieces of evidence in constructing moral theories. Rival moral principles, such as those constituting deontology and consequentialism, are tested by seeing if they get it right on certain important cases.

Take the well-known example of the Trolley Problem. An out-of-control trolley is rumbling toward several innocents trapped on a track. You can divert the trolley, but only by sending it onto a side track where it will kill one person. Most people think it is morally permissible to divert the trolley in this case. Now imagine that the trolley cannot be diverted—it can be stopped only by physically pushing a large person into its path, causing an early crash. Most people do not think this is morally permitted. If we treat these two intuitive reactions as moral evidence, then we can affirm that how an individual is killed makes a difference to the rightfulness of sacrificing a smaller number of lives to save a larger number. Apparently, it is permissible to indirectly sacrifice an innocent as a side-effect of diverting a threat, but not permissible to directly use an innocent as a means to stop the threat. This seems like preliminary evidence against a moral theory that says that outcomes are the only things that matter morally, since the outcomes are identical in these two cases.

This sort of reasoning is at the center of how most philosophers practice normative ethics. Moral principles are proposed to account for intuitive reactions to cases. The best principles are (all else equal) those that cohere with the largest number of important intuitions. A philosopher who wishes to challenge a principle will construct a clever counterexample: a case where it just seems obvious that it is wrong to do X, but the targeted principle allows us to do X in the case. Proponents of the principle must now (a) show that their principle has been misapplied and actually gives a different verdict about the case; (b) accept that the principle has gone wrong but alter it to give a better answer; (c) bite the bullet and insist that even if the principle seems to have gone wrong here, it still trustworthy because it is right in so many other cases; or (d) explain away the problematic intuition, by showing that the test case is underdescribed or somehow unfair, or that
the intuitive reaction itself likely results from confusion. If all of this sounds a bit like the testing of scientific hypotheses against experimental data, that is no accident. The philosopher John Rawls (1971) explicitly modeled this approach, which he called “reflective equilibrium,” See “Rawls” on hypothesis testing in science.

There are various ways to understand what reflective equilibrium aims at doing. In a widely accepted interpretation, reflective equilibrium aims at discovering the substantive truths of ethics. In this understanding, those moral principles supported in reflective equilibrium are the ones most likely to be the moral truth. (We will leave aside here How to interpret truth in the moral domain is left aside in this article, but see “Metaethics.”) Intuitions about test cases are evidence for moral truth in much the same way that scientific observations are evidence for empirical truth. In science, our confidence in a particular theory depends on whether it gets evidential support from repeated observations, and in moral philosophy (in this conception) our confidence in a particular ethical theory depends on whether it gets evidential confirmation from multiple intuitions.

This parallel allows us to see one way in which the psychology of moral judgment might be relevant to moral philosophy. When we are testing a scientific theory, our trust in any experimental observation depends on our confidence in the scientific instruments used to generate them. If we come to doubt the reliability of our instruments, then we should doubt the observations we get from them, and so should doubt the theories they appear to support. What, then, if we come to doubt the reliability of our instruments in moral philosophy? Of course, philosophers do not use microscopes or mass spectrometers. Our instruments are nothing more than our own minds—or, more precisely, our mental abilities to understand situations and apply moral concepts to them. Could our moral minds be unreliable in the way that microscopes can be unreliable?

This is certainly not an idle worry, since we know that we make consistent mental mistakes in other domains. Think about persistent optical illusions or predictably bad mathematical reasoning. There is an enormous literature in cognitive science showing that we make consistent mistakes in domains other than morality. Against this background, it would be a surprise if our moral intuitions turned out not to be full of mistakes.

In earlier sections we discussed psychological evidence showing systematic deficits in moral reasoning. We saw, for instance, that people’s moral judgments are affected by the verbal framing in which test cases are presented (save versus die) and by the cleanliness of their immediate environment. If the readings of a scientific instrument appeared to be affected by environmental factors that had nothing to do with what the instrument was supposed to measure, then we would rightly doubt the trustworthiness of readings obtained from that instrument. In moral philosophy, a mere difference in verbal framing, or the dirtiness of the desk you are now sitting at, certainly do not seem like things that matter to whether or not a particular act is permissible. So it seems that our moral intuitions, like defective lab equipment, sometimes cannot be trusted.
Note how this argument relates to earlier claims about the autonomy of morality or the is/ought distinction. Here no one is claiming that cognitive science tells us directly which moral judgments are right and which wrong. All cognitive science can do is show us that particular intuitions are affected by certain causal factors—it cannot tell us which causal factors count as distorting and which are acceptable. It is up to us, as moral judges, to determine that differences of verbal framing (save/die) or desk cleanliness do not lead to genuine moral differences. Of course, we do not have to think very hard to decide that these causal factors are morally irrelevant—but the point remains that we are still making moral judgments, even very easy ones, and not getting our verdict directly from cognitive science.

Many proponents of a role for cognitive science in morality are willing to concede this much: in the end, any revision of our moral judgments will be authorized only by some other moral judgments, not by the science itself. But, they will now point out, there are some revisions of our moral judgments we ought to make, and we are able to make only because of input from cognitive science. We all agree that our moral judgments should not be affected by the cleanliness of our environment, and the science is unnecessary for our agreeing on this. But we would not know about the fact that our moral judgments are affected by environmental cleanliness without cognitive science. So, in this sense at least, improving the quality of our moral judgments does seem to require the use of cognitive science. Put more positively: paying attention to the cognitive science of morality can allow us to realize that some seemingly reliable intuitions are not in fact reliable. Once these are set aside like broken microscopes, we can have greater confidence that the theories we build from the remainder will capture the moral truth. (See, for example, Sinnott-Armstrong 2008; Mason 2011.)

This argument is one way of showing the relevance of cognitive science to morality. Note that it is an epistemic debunking argument. The argument undermines certain moral intuitions as default sources of evidential justification. Cognitive science plays a debunking role by exposing the unreliable causal origins of our intuitions. Philosophers who employ debunking arguments like this do so with a range of aims. Some psychological debunking arguments are narrowly targeted, trying to show that a few particular moral intuitions are unjustified. Often this is meant to be a move within normative theory, weakening a set of principles the philosopher rejects. For instance, if intuitions supporting deontological moral theory are undermined, then we may end up dismissing deontology. Other times, philosophers intend to aim psychological debunking arguments much more widely. If it can be shown that all moral intuitions are undermined in this way, then we have grounds for skepticism about moral judgment [see “Moral Epistemology.”] The plausibility of all these arguments remains hotly debated. But if any of them should turn out to be convincing, then we have a clear demonstration of how cognitive science can matter to moral theory.

9. Non-epistemic Approaches
Epistemic debunking arguments presuppose that morality is best understood as an epistemic domain. That is, these arguments assume that there is (or could be) a set of moral truths, and that the aim of moral judgment is to track these moral truths. What if we do not accept this conception of the moral domain? What if we do not expect there to be moral truths, or do not think that moral intuition aims at tracking any such truth? In that case, should we care about the cognitive science of morality?

**a. Consistency in Moral Reasoning**

Obviously the answer to this question will depend on what we think the moral domain is, if it is not an epistemic domain. One common view, often associated with contemporary Kantians, is that morality concerns *consistency in practical reasoning*. Though we do not aim to uncover independent moral truths, we can still say that some moral beliefs are better than others, because some moral beliefs are better at cohering with the way in which we conceive of ourselves, or what we think we have most reason to do. In this understanding, a bad moral belief is not bad because it is mistaken about the moral facts (there are no moral facts) but is bad because it does not fit well with our other normative beliefs. The assumption here is that we want to be coherent, in that being a rational agent means aiming at consistency in the beliefs that ground one’s actions. [See “Moral Epistemology.”]

In this conception of the moral domain, cognitive science is useful because it can show us when we have unwittingly stumbled into inconsistency in our moral beliefs. A very simple example comes from the *universalizability condition* on moral judgment. It is incoherent to render different moral judgments about two people doing the exact same thing in the exact same circumstances. This is because morality (unlike taste, for instance) aims at universal prescription. If it is wrong for you to steal from me, then it is wrong for me to steal from you. (Assuming you and I are similarly situated—if I am an unjustly rich oligarch and you are a starving orphan, maybe things are different. But then this moral difference is due to different features of our respective situations. The point of universal prescription is to deny that there can be moral differences when situations are similar.) A person who explicitly denied that moral requirements applied equally to herself as to other people would not seem to really have gotten the point of morality. We would not take such a person very seriously when she complained of others violating her moral rights, if she claimed to be able to ignore those same rights in others.

We all know that we fail at universalizing our moral judgments sometimes; we all suffer moments of weakness where we try to make excuses for our own moral failings, excuses we would not permit to others. But some psychological research suggests that we may fail in this way far more often than we realize. Nadelhoffer and Feltz (2008) found that people make different judgments about moral dilemmas when they imagine themselves in the dilemma than when imagining others in the same dilemma. Presumably most people would not explicitly agree that there is such a moral
difference, but they can be led into endorsing differing standards depending on whether they are presented with a me versus someone else framing of the case. This is an unconscious failure of universalization, but it is still an inconsistency. If we aim at being consistent in our practical reasoning, we should want to be alerted to unconscious inconsistencies, so that we might get a start on correcting them. And in cases like this one, we do not have introspective access to the fact that we are inconsistent, but we can learn it from cognitive science.

Note how this argument parallels the epistemic one. The claim is, again, not that cognitive science tells us what counts as a good moral judgment. Rather cognitive science reveals to us features of the moral judgments we make, and we must then use moral reasoning to decide whether these features are problematic. Here the claim is that inconsistency in moral judgments is bad, because it undermines our aim to be coherent rational agents. We do not get that claim from cognitive science, but there are some cases where we could not apply it without the self-knowledge we gain from cognitive science. Hence the relevance of cognitive science to morality as aimed at consistency.

b. Rational Agency

There is another way in which cognitive science can matter to the coherent rational agency conception of morality. Some findings in cognitive science may threaten the intelligibility of this conception altogether. Recall, from section 2, the psychologist Jonathan Haidt’s work on moral dumbfounding; people appear to spontaneously invent justifications for their intuitive moral verdicts, and stick with these verdicts even after the justifications are shown to fail. If pressed hard enough, people will admit they simply do not know why they came to the verdicts, but hold to them nevertheless. In Haidt’s interpretation, these findings show that moral judgment happens almost entirely unconsciously, with conscious moral reasoning mostly a post hoc epiphenomenon.

If Haidt is right, point out Jeanette Kennett and Cordelia Fine (2009), then this poses a serious problem for the ideal of moral agency. For us to count as moral agents, there needs to be the right sort of connection between our conscious reasoning and our responses to the world. A robot or a simple animal can react, but a rational agent is one that can critically reflect upon her reasons for action and come to a deliberative conclusion about what she ought to do. Yet if we are morally dumbfounded in the way Haidt suggests, then our conscious moral reasoning may lack the appropriate connection to our moral reactions. We think that we know why we judge and act as we do, but actually the reasons we consciously endorse are mere post hoc confabulations.

In the end, Kennett and Fine argue that Haidt’s findings do not actually lead to this unwelcome conclusion. They suggest that he has misinterpreted what the experiments show, and that there is a more plausible interpretation that preserves the possibility of conscious moral agency. Note that responding in this way concedes that cognitive science might be relevant to assessing the status of
our moral judgments. The dispute here is only over what the experiments show, not over what the implications would be if they showed a particular thing. This leaves the door open for further empirical research on conscious moral agency.

One possible approach is the selective application of Haidt’s argument. If it could be shown that certain moral judgments—those about a particular topic or sub-domain of morality—are especially prone to moral dumbfounding, then we might have the basis for disqualifying them from inclusion in reflective moral theory. This seems, at times, to be the approach adopted by Joshua Greene (see section 4) in his psychological attack on deontology. According to Greene (2014), deontological intuitions are of a psychological type distinctively disconnected from conscious reflection and should accordingly be distrusted. Many philosophers dispute Greene’s claims (see, for instance, Kahane 2012), but this debate itself shows the richness of engagement between ethics and cognitive science.

c. Intersubjective Justification

There is one further way in which cognitive science may have relevance to moral theory. In this last conception, morality is essentially concerned with intersubjective justification. Rather than trying to discover independent moral truths, my moral judgments aim at determining when and how my preferences can be seen as reasonable by other people. A defective moral judgment, in this conception, is one that reflects only my own personal idiosyncrasies and so will not be acceptable to others. For instance, perhaps I have an intuitive negative reaction to people who dress their dogs in sweaters even when it is not cold. If I come to appreciate that my revulsion of this practice is not widely shared, and that other people cannot see any justification for it, then I may conclude that it is not properly a moral judgment at all. It may be a matter of personal taste, but it cannot be a moral judgment if it has no chance of being intersubjectively justified.

Sometimes we can discover introspectively that our putative moral judgments are actually not intersubjectively justifiable, just by thinking carefully about what justifications we can or cannot offer. But there may be other instances in which we cannot discover this introspectively, and where cognitive science may help. This is especially so when we have unknowingly confabulated plausible-sounding justifications in order to make our preferences appear more compelling than they are (Rini 2013). For example, suppose that I have come to believe that a particular charity is the most deserving of our donations, and I am now trying to convince you to agree. You point out that other charities seem to be at least as effective, but I insist. By coincidence, the next day I participate in a psychological study of color association. The psychologists’ instruments notice that I react very positively to the color forest green—and then I remember that my favorite charity’s logo is a deep forest green. If this turns out to be the explanation for why I argued for the charity, then I should doubt that I have provided an intersubjective justification. Maybe my fondness for the charity’s logo is an okay reason for me to make a personal choice (if I am otherwise indifferent),
but it certainly is not a reason for you to agree. Now that I am aware of this psychological influence, I should consider the possibility that I have merely confabulated the reasons I offered to you.

10. Objections and Alternatives

The preceding sections have focused on negative implications of the cognitive science of moral judgment. We have seen ways in which learning about a particular moral judgment’s psychological origins might lead us to disqualify it, or at least reduce our confidence in it. This final section briefly considers some objections to drawing such negative implications, and also discusses more positive proposals for the relationship between cognitive science and moral philosophy.

a. Explanation and Justification

One objection to disqualifying a moral judgment on cognitive scientific grounds is that this involves confusion between reasons of explanation and justification. The explanatory reason for the fact that I judge X to be immoral could be any number of psychological factors. But my justifying reason for the judgment is unlikely to be identical with the explanatory reason. Consider my judgment that it is wrong to tease dogs with treats that will not be provided. Perhaps the explanatory reason for my believing this is that my childhood dog bit me when I refused to share a sandwich. But this is not how I justify my judgment—the justifying reason I have is that dogs suffer when led to form unfulfilled expectations, and the suffering of animals is a moral bad. As long as this is a good justifying reason, then the explanatory reason does not really matter. So, runs the objection, those who disqualify moral judgments on cognitive scientific grounds are looking at the wrong thing—they should be asking about whether the judgment is justified, not why (psychologically speaking) it was made (Kamm 1998; van Roojen 1999).

One problem with this objection is that it assumes we have a basis for affirming the justifying reasons for a judgment that are unaffected by cognitive scientific investigation. Obviously if we had oracular certainty that judgment X is correct, then we should not worry about how we came to make the judgment. But in moral theory we rarely (if ever) have such certainty. As discussed earlier (see section 8), our justification for trusting a particular judgment is often dependent upon how well it coheres with other judgments and general principles. So if a cognitive scientific finding showed that some dubious psychological process is responsible for many of our moral judgments, their ability to justify one another may be in question. To see the point, consider the maximal case: suppose you learned that all of your moral judgments were affected by a chemical placed in the water supply by the government. Would this knowledge not give you some reason to second-guess your moral judgments? If that is right, then it seems that our justifying reasons for holding to a judgment can be sensitive to at least some discoveries about the explanatory reasons for them. (For related arguments, see Street 2006 and Joyce 2006.)
b. The Expertise Defense

Another objection claims to protect the judgments used in moral theory-making even while allowing the in-principle relevance of cognitive scientific findings. The claim is this: cognitive science uses research subjects who are not experts in making moral judgments. But moral philosophers have years of training at drawing careful distinctions, and also typically have much more time than research subjects to think carefully about their judgments. So even if ordinary participants in cognitive science studies make mistakes due to psychological quirks, we should not assume that the judgments of experts will resemble those of non-experts. We do not doubt the competence of expert mathematicians simply because the rest of us make arithmetic mistakes (Ludwig 2007). So, the objection runs, if it is plausible to think of moral philosophers as experts, then moral philosophers can continue to rely upon their judgments whatever the cognitive science says about the judgments of non-experts.

Is this expertise defense plausible? One major problem is that it does not appear to be well supported by empirical evidence. In a few studies (Schwitzgebel and Cushman 2012; Tobia, Buckwalter, and Stich 2013), people with doctorates in moral philosophy have been subjected to the same psychological tests as non-expert subjects and appear to make similar mistakes. There is some dispute about how to interpret these studies (Rini 2015), but if they hold up then it will be hard to defend the moral judgments of philosophers on grounds of expertise.

c. The Regress Challenge

A final objection comes in the form of a regress challenge. Henry Sidgwick first made the point, in his *Methods of Ethics*, that it would be self-defeating to attempt to debunk judgments on the grounds of their causal origins. The debunking itself would rely on some judgments for its plausibility, and we would then be led down an infinite regress in querying the causal origins of these judgments, and causal origins of the judgments responsible for our judgments about those first origins, and so on. Sidgwick seems to be discussing general moral skepticism, but a variant of this argument presents a regress challenge to even selective cognitive scientific debunking of particular moral judgments. According to the objection, once we have opened the door to debunking, we will be drawn into an inescapable spiral of producing and challenging judgments about the moral trustworthiness of various causal origins. Perhaps, then, we should not start on this project at all.

This objection is limited in effect; it applies most obviously to epistemic forms of cognitive scientific debunking. In non-epistemic conceptions of the aims of moral judgment, it may be possible to ignore some of the objection’s force. The objection is also dependent upon certain empirical assumptions about the interdependence of the causal origins driving various moral judgments. But if sustained, the regress challenge for epistemic debunking seems significant.
d. Positive Alternatives

Finally, we might consider an alternative take on the relationship between moral judgment and cognitive science. Unlike most of the approaches discussed above, this one is positive rather than negative. The idea is this: if cognitive science can reveal to us that we already unconsciously accept certain moral principles, and if these fit with the judgments we think we have good reason to continue to hold, then cognitive science may be able to contribute to the construction of moral theory. Cognitive science might help you to explicitly articulate a moral principle that you already accepted implicitly (Mikhail 2011; Kahane 2013). In a sense, this is simply scientific assistance to the traditional philosophical project of making explicit the moral commitments we already hold—the method of reflective equilibrium developed by Rawls and employed by most contemporary ethicists. In this view, the use of cognitive science is likely to be less revolutionary, but still quite important. Though negative approaches have received most discussion, the positive approach seems to be an interesting direction for future research.

11. References and Further Reading


Regina A. Rini
Email: gina.rini@nyu.edu
New York University
U. S. A.