Moral Rationalism on the Brain

Joshua May

Published in *Mind & Language* 38(1): 237-255
DOI: 10.1111/mila.12394

Abstract: I draw on neurobiological evidence to defend the rationalist thesis that moral judgments are essentially dependent on reasoning, not emotions (conceived as distinct from inference). The neuroscience reveals that moral cognition arises from domain-general capacities in the brain for inferring, in particular, the consequences of an agent's action, the agent's intent, and the rules or norms relevant to the context. Although these capacities entangle inference and affect, blurring the reason/emotion dichotomy doesn't preferentially support sentimentalism. The argument requires careful consideration of the empirical evidence (from neuroimaging to psychopathology) and philosophical analysis of the commitments of rationalism versus sentimentalism in ethics.

Keywords: moral reasoning, moral emotion, psychopathy, moral dilemmas, acquired sociopathy

1. Introduction

Moral rationalism is the thesis that ethics is grounded in reason, but the tradition is varied. It is typically thought of as a package of views, including the claims that moral judgments are reasoned and that moral facts are, like mathematical truths, knowable *a priori* through pure understanding or conceptual competence (Jones & Schroeter 2018). My topic is only the psychological thesis about moral judgment and how the burgeoning field of moral neuroscience bears on it.

Most philosophers and scientists appear to take empirical evidence to preferentially support the alternative view, moral sentimentalism, associated with the likes of Adam Smith and David Hume, among others. According to a traditional brand of sentimentalism, emotions (conceived as distinct from reason) are essential and fundamental to distinctively moral cognition. Experimental research seems to suggest that emotions amplify moral judgment and that moral incompetence (e.g. in psychopaths) involves emotional deficits. Jesse Prinz (2016: 45-6), for example, interprets "the preponderance of empirical evidence as supporting a fairly traditional kind of sentimentalist theory of moral judgment" (see also Haidt 2001; Nichols 2004; Sinhababu 2017; Kauppinen 2019).

Some rationalists have resisted this trend by drawing on work in psychological science but little on neurobiology, if at all (Maibom 2005; Kennett 2006; Horgan & Timmons 2007; Sauer 2017; May 2018a). Perhaps that is because philosophers typically assume that neuroscience—which only examines the brain's hardware, so to speak—is unable to illuminate the psychological mechanisms and ethical issues (Berker 2009).

Here I aim to support rationalism by drawing on a broad range of neurobiological evidence. I show that the neuroscience, when combined with relevant psychological science and philosophical analysis, adds positive evidence that corroborates, illuminates, and strengthens the empirical case for moral rationalism. The argument proceeds in two steps. First, although inference and affect appear to be entangled in the brain, rationalism is surprisingly well-suited to explain this blurring of the reason/emotion dichotomy (§3). Second, moral cognition appears to arise from domain-general capacities in the brain, such as inferring the consequences of an agent's action and the agent's intent (§4). I conclude that rationalism is not only consistent with the resulting picture of the moral brain but provides a rather satisfying account of it (§5). Analyzing the empirical evidence, however, requires first clarifying the commitments of the two camps (§2).

2. Rationalism vs. Sentimentalism in Ethics

2.1 The Classical Theories

Rationalism and sentimentalism are such broad traditions in ethics that any characterization of them will inevitably be unsatisfying to some parties in the debate. Recognizing these constraints, we'll consider some traditional forms of these theories with the acknowledgement that they can be modified and updated in light of improved understandings of the human mind. Indeed, the aim here is to provide a version of rationalism that coheres with moral neuroscience, even if updated versions of sentimentalism are available. We begin with such classical formulations of the theories.

Sentimentalists have traditionally suggested that moral judgments are like judgments of beauty in requiring certain emotional responses (Gill 2007). You don't need any emotional reactions to the thought that "Tehran is the capital of Iran" in order to accept it, and the same goes for the mathematical thought that "To infer the length of a right triangle's hypotenuse from the length of its other sides, one ought to use the Pythagorean theorem." But it's hard to imagine truly believing that "Yosemite valley is breathtakingly beautiful" without ever having that distinctive feeling of awe that it inspires. Similarly, without feelings of compassion or indignation, sentimentalists think it impossible to truly make particular moral judgments about helping others in need or rectifying social injustices. Some sentimentalists go so far as to say that moral judgments are "constituted by emotional states" (Prinz 2016: 46; see also D'Arms & Jacboson 2014: 254; Tappolet 2016: 79). At a minimum, however, defenders of the psychological thesis of *sentimentalism* traditionally contend that moral cognition is fundamentally caused by such emotional responses (Nichols 2004: 83; see also Kauppinen 2018).

In contrast, proponents of *rationalism* traditionally define the view in terms of reasoning and rationality. Moral judgment, it is said, fundamentally "derives from our rational capacities" (Kennett 2006: 70), is ultimately "the culmination of a process of reasoning" (Maibom 2010: 999), or is "fundamentally an inferential enterprise that is not ultimately dependent on non-rational emotions, sentiments, or passions" (May 2018a: 7). Emotions, according to rationalists, are either merely the natural consequences of reasoning or provide just one way of instigating or facilitating reasoning or inference, which of course can be downright irrational while being part of our "rational capacities."

These characterizations of the traditional debate might seem to rely on a sharp reason/emotion dichotomy, which has been duly questioned (more on this below). However, we can preserve a classical dispute by asking whether the affective elements of moral emotions are necessary for moral judgment (or, more broadly, normative/evaluative judgments about what one should do). Sentimentalists maintain that moral judgments are *special* in requiring emotions in a way that other forms of cognition don't, such as judgments about mathematics, geography, or language. Consider how the issue is discussed by Hume. He says certain topics are the objects of reason while others the objects of feeling. To make the point that morality is special in being an object of feeling, he asks us to consider any vicious action and to look for a "matter of fact" that makes it vicious:

> The vice entirely escapes you, as long as you consider the object [the person with the vice]. You never can find it, till you turn your reflection into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action. Here is a matter of fact; but 'tis the object of feeling, not of reason. (1739–40/2000: 3.1.1)

Of course, as Hume emphasizes elsewhere, reasoning is involved in making such moral judgments. But sentimentalists will maintain that it's the feelings, the affective elements of moral emotions, that are essential for and distinctive of moral cognition. Sentimentalists have not traditionally argued that feelings are required for understanding right from wrong because feelings are required for *all* knowledge. Put another way, imagine a company aims to build military androids or autonomous vehicles with knowledge of mathematics but also of ethics, so that the AI can properly navigate moral dilemmas and avoid atrocities. Or imagine a science-fiction writer who is trying to construct a plausible fictional species from a distant galaxy who knows right from wrong and yet has a mind that differs in other respects from human psychology. Although the machines or fictional species might be able to make moral judgments of a sort, sentimentalists typically regard the resulting judgments as derivative at best. Their predicament would be analogous to how a congenitally blind person can judge that ripe bananas are yellow by relying on the visual experiences of others (Prinz 2006: 32). Unlike knowledge of mathematics, ethical knowledge is ultimately grounded in moral emotions (e.g. compassion, indignation), whether experienced by oneself or others.

Rationalists, on the other hand, would insist on treating both kinds of knowledge similarly. While ethics is a distinct domain, it too deploys *domain-general mechanisms* of reasoning or inference, whether conscious or unconscious. An android must of course possess moral concepts in order to make moral judgments—just as it would have to possess mathematical concepts, like *addition* and *hypotenuse*, in order to make mathematical judgments—but emotions aren't necessary for possessing the concepts. Rationalists of course will admit that humans are emotional creatures and that these emotions help to quickly draw our attention to morally relevant information. But their view implies that a creature with unlimited time and resources needn't possess emotions in order to make distinctively *moral* judgments (cf. Jones 2006: 3). Indeed, rationalists can and should predict that *morality isn't special*: emotions, affect, and feelings do aid reasoning generally, across a wide range of subject matters or domains of thought, from morality to mathematics. That prediction, I'll argue, looks to be supported by our best understanding of moral cognition in the human brain.

2.2 The Terms

Some important clarifications are in order. First, although the paradigm of reasoning is conscious deliberation, we must resist the urge to identify reasoning with this narrow class of mental phenomena. Otherwise we unfairly saddle rationalists with the daft thesis that moral judgment is fundamentally driven by slow, conscious deliberation that is entirely bereft of feeling (May & Kumar 2018). Instead, we will charitably interpret the rationalist tradition as maintaining that moral judgment is fundamentally a matter of *inference*. Although inference needn't be conscious, it is more than mere information processing or computation. Inference is commonly understood as the process of forming beliefs (or belief-like states) on the basis of other such states, whether consciously or unconsciously (Boghossian 2012; Siegel 2017: ch. 5; Buckner 2019). For example, you conclude (form the belief) that your date is trustworthy on the basis of your perceptions of his demeanor, affectations, and relevant behavior toward you and others. A complete analysis of "on the basis of" has proven elusive. Inference requires more than your existing attitudes merely causing you to form a new belief through, say, mere priming or association. But must you believe or recognize that your conclusion is supported by the evidence? We needn't provide a definitive answer here, but instead only assume what is uncontroversial: that conscious recognition isn't required. Otherwise too many genuine instances of inference are ruled out.

Of course, moral judgment involves inference applied to decision and action, not merely abstract theoretical reasoning. An android would probably be incapable of full-fledged moral judgment if it is merely imbued with reasoning and lacks any behavior, goals, or desires. However, inference mediates not only prospective deliberation but unconscious learning. Rationalists would expect the android to acquire moral judgment if we put it in an environment with other social creatures along with domain-general learning capacities, including theory of mind, calculation of outcomes, and the ability to categorize act types as being consistent with or violating general rules or norms. Rather than model moral cognition on reasoning through a mathematical proof, modern rationalists take a page from recent advances in artificial intelligence that use deep learning algorithms to model human cognition, in ethics and elsewhere (Malle & Scheutz 2019; Haas 2020).

Second, the form of rationalism I defend is non-partisan regarding neighboring issues about the reliability or rationality of moral judgments. Like the android, humans have goals and desires, which can lead to wishful thinking, confirmation bias, and post-hoc rationalization. Yet problematic forms of inference are inference nonetheless. Thus, the neuroscience reveals a role for reasoning that is compatible with both attempts to debunk (Greene 2014) and to vindicate (Kumar & May 2019) commonsense moral intuitions.

Finally, amid the replication crisis in science, we shouldn't be overly credulous when drawing on scientific research, whether the discipline is a natural or social science (Machery & Doris 2017; May 2021). We'll see that the argument here is based on more than a small set of studies or methods from a single lab. Moreover, the research involves more than brain imaging (and problematic "reverse inferences") or moral dilemmas, such as the trolley problem, which might be too artificial or unrealistic to serve as the sole measure of everyday moral cognition.

3. Step 1: The Entanglement of Inference and Affect

Early work in moral neuroscience seemed to reveal the centrality of emotion, a point emphasized by some sentimentalists (e.g. Prinz 2016). However, as we'll see, this work also calls into question the common distinction between reason and emotion, revealing that apparently "emotional" processes are important not just for moral cognition but for reasoning generally, outside of the moral domain. This result is not only consistent with the tradition of rationalism but predicted by it and in conflict with classical forms of sentimentalism. The entanglement of inference and affect has been most exemplified in neurological disorders and brain damage. Here I focus on two key cases: psychopathy and so-called "acquired sociopathy."

3.1 Psychopathy

First consider how both affective and rational deficits are entangled in the abnormal moral thought and behavior of psychopaths. Psychopathy (sometimes referred to as "sociopathy") is roughly what's labeled "anti-social personality disorder" in the latest Diagnostic and Statistical Manual of Mental Disorders (DSM-5, American Psychological Association, 2013). Among other vices, psychopaths are characteristically callous, shameless, remorseless, and exhibit a superficial charm that allows them to manipulate others who are often ultimately left battered or destitute. These symptoms do appear to contribute to some misunderstanding of right and wrong. Some, though not all, inmates with psychopathy have difficulty distinguishing moral rules from mere conventions (Aharoni et al. 2012); others struggle to properly deploy moral concepts and reasons in conversation—e.g., misusing words like “guilt” and “feeling bad” about a crime (Hare 1993; Kennett & Fine 2008).

Most psychopaths aren’t serial killers, although they commonly have run-ins with the law. Hare (1993: 91) describes one case, for example, in which a man with psychopathy broke into what he thought was an empty home but found an irate resident inside who wouldn’t “shut up.” Instead of fleeing, the burglar calmly beat the elderly man unconscious, then took a nap on the sofa, only to be woken up later by police. Such senseless aggression and instrumental irrationality are common in psychopathy (Maibom 2005).

The disorder seems to arise primarily from dysfunction in at least two key areas of the brain and their connectivity (Blair 2007; Kiehl 2006). First is the *ventromedial prefrontal cortex* (vmPFC), which is roughly the inner underside of the “prefrontal” cortex (the front of the frontal lobe). The vmPFC overlaps with a similar anatomical region behind the eye sockets (or “orbits”), the orbitofrontal cortex. The frontal lobe generally subserves our ability to plan and make complex choices, and these specific portions of the prefrontal cortex behind the eyes seem to house the gut feelings that guide judgment and decision-making (more on this in the next section). The second main area of dysfunction in psychopathy is the *amygdala*, a pair of small almond-shaped nodes deep in the brain. These nodes, along with the limbic system of which they’re a part, have classically been associated with emotion, particularly fear. But we now know they are involved more broadly in assessing the significance of an object or event—e.g., whether it’s valuable or threatening. The amygdala is thus a region crucial for many processes of learning in light of rewards or punishments, and dysfunction early in development naturally leads to a host of cognitive and behavioral problems. Proper development of the vmPFC and amygdala is never fully achieved for individuals with psychopathy, for a variety of reasons. Part of the explanation is genetic, such as mutations that disrupt neurotransmitters, but other important factors include adverse circumstances like neglect, childhood trauma, and even exposure to lead (Kiehl & Sinnott-Armstrong 2013; Glenn & Raine 2014).

Many theorists have remarked that psychopaths seem to be in some sense rational, at least in their general understanding of the world and how to go about manipulating it. A common analysis is that psychopaths thus demonstrate the necessity of emotions for understanding right from wrong (e.g. Haidt 2001; Nichols 2004; Prinz 2016). However, psychopathy exhibits the entanglement of inference and affect in moral cognition. Deficits in feeling guilt and compassion are accompanied by numerous rational deficits—poor attention span, delusions of grandeur, and difficulty learning from punishment (Maibom 2005; May 2018a). The impairments are broadly in learning and inference, commonly experienced early in life, which prevent patients from properly understanding and internalizing habits and strategies for ethical and prudential decision-making. Indeed, as we’ll see in the next section, adults who acquire damage to areas like the vmPFC later in life, after such areas have already developed properly, display markedly different psychological profiles (Anderson et al. 1999; Taber-Thomas et al. 2014).

3.2 “Acquired Sociopathy”

A similar demonstration of the entanglement of inference and affect arises from lesions of the vmPFC in adults. This area of the prefrontal cortex appears to be roughly the site of damage in Phineas Gage, the railroad worker whose tamping iron rocketed through his frontal lobe in 1848.

The details of that famous case are fraught, but we know much more now about individuals who have suffered damage to this area in adulthood.

Patients with vmPFC damage often develop what Antonio Damasio (1994) dubs "acquired sociopathy." This label is confusing, however, since the clinical profile is quite different from that of psychopathy. Far from callous or remorseless, adults who acquire damage to the vmPFC primarily suffer from poor decision-making that is not necessarily antisocial or harmful to others. These patients appear instead to have a shortage of gut feelings that help guide decisions about what to do in the moment. Compared to neurotypical individuals, patients have difficulty making a wide range of decisions, from how to rack up points in a card game to which variety of apples to purchase at the grocery store. Consider two patients as examples:

- Elliot, one of Damasio's (1994: ch. 3) patients, had an orange-sized tumor that originated near his nasal cavity. Removal of the tumor resulted in damage to his frontal lobe (primarily orbital and medial portions). After surgery, Elliot remained intelligent in many respects but struggled with all manner of planning and decision-making. He was eventually fired from his job because he could no longer stay on task and properly manage his time. His relationships suffered as well from what some described as his foolish and irrational behavior, leading to multiple marriages and divorce.
- Tammy Myers presents a similar form of decision-making deficit. After a motorcycle accident damaged her orbitofrontal cortex, Tammy reports that "she often spends all day on the sofa" because even simple decisions about what to do next are agonizing (Eagleman 2015: 119).

Patients with such brain damage retain many of their intellectual capacities. They tend to give typical responses about how one ought to make hypothetical choices (Saver & Damasio 1991). The deficit rather concerns decisions about what to do *oneself* in a *particular situation* (Kennett & Fine 2008). A patient might recognize that it's healthier to buy organic fruits, but what should *she* do right *now* when the non-organic Opal apples this season taste divine? Although similar decision paralysis crops up in social situations too, the deficit is domain-general.

These decision-making deficits appear to affect patients' rationality. But Damasio (1994) and many others following him attribute the deficit in decision-making to "somatic markers" that help us to settle on decisions that feel right. The absence or attenuation of a patient's gut feelings thus impairs one's ability to choose among competing options. However, as Damasio emphasizes, acquired sociopathy exemplifies the entanglement of inference and affect in decision-making generally.

Indeed, later research suggests that the vmPFC is itself unlikely a source of gut feelings but rather a hub wherein such feelings are incorporated with or weighed against other considerations to then form a decision (Shenhav & Greene 2014; Hutcherson et al. 2015; Salzman & Fusi 2010; Woodward 2016). Other lines of research suggest that the amygdala and vmPFC in mammals facilitate certain forms of reinforcement and statistical learning that are largely unconscious (e.g., Hare et al. 2008) and can be applied to moral judgment among other domains (Cushman 2013; Crockett 2016; Nichols 2021). It's no wonder that patients with damage to the vmPFC struggle to make decisions. It's not because they lack knowledge of right and wrong or are unable to form moral judgments whatsoever; rather absent or attenuated gut feelings impair their ability integrate their values into an overall decision about what to do right now in these particular circumstances.

3.3 Entanglement & Rationalism

If the reason/emotion dichotomy is blurry, then this might seem to trivialize the debate between rationalists and sentimentalists. However, the entanglement of inference and affect doesn't support the traditional sentimentalist thesis that, unlike other forms of judgment, moral judgment requires *moral emotions* such as anger, compassion, guilt, and disgust. Rather, the emerging picture is that twinges of affect are necessary for all forms of cognition, from judgments about ethics to judgments about which apples to purchase (Huebner 2015; Seligman et al. 2016). Indeed, a growing consensus in neurobiology is that "affect is a form of cognition" (Duncan & Barrett 2007), such that even if, say, the amygdala gives rise to affect, such feelings facilitate inference generally. The consensus comports with the classical rationalist thesis that moral cognition isn't special; it involves affect in just the way that other forms of inference do. It's perfectly consistent with rationalism to hypothesize that emotions draw one's attention to morally relevant information, as some sentimentalists have recognized (e.g. Prinz 2006: 31). The debate might break down if contemporary sentimentalists abandon the idea that moral or evaluative judgment is special in requiring emotions. Then both camps might happily embrace the entanglement of inference and affect, but a dispute remains intelligible between more traditional views.

Classical sentimentalists might argue that the amygdala and vmPFC provide the crucial affective or emotional element that makes for a distinctively moral judgment, or more broadly normative/evaluative judgments about what one should or shouldn't do. But that's not quite right, for several reasons.

First, if we focus on full-blown moral emotions like compassion and moral disgust, the reason/emotion dichotomy isn't so blurry. Experimental and anecdotal evidence suggests that such rich moral emotions are commonly the effects, not causes, of moral judgments. Often one only feels compassion for a group of people, such as immigrants, if one antecedently judges them to be unfairly disadvantaged or deserving of sympathy (Betancourt 1990). And we feel repugnance toward those acts we antecedently deem immoral (Plakias 2018; May 2018b). In a set of electroencephalography (EEG) studies, for example, carefully time-locked recordings of participants' brain waves suggest that they first judge an action to be moral or immoral and then, milliseconds later, judge the act to be disgusting or not (Yang et al 2013; 2014). Of course, emotions can be recalcitrant in the face of changes in moral judgments—e.g., one might not become disgusted by the consumption of animal flesh, even after judging it to be immoral. Yet people who become vegetarians for moral, as opposed to health, reasons are apparently more likely to become disgusted by meat (Rozin et al. 1997). In this way, emotions often fall in line with the new moral belief, as when anger toward your partner eventually fades after realizing her remarks weren't disrespectful. So, even if one's moral judgments arise from learning and inference that is suffused with affect, this entanglement is no consolation for classical sentimentalism when there remains a dichotomy between the reasoned judgment and the full-blown moral emotions that result. In other words, although the science militates against a dichotomy between inference and affect, the evidence speaks in favor of distinguishing to some degree reasoned moral judgments from moral emotions (and in a way that is compatible with a rationalist model).

Sentimentalists might focus less on distinctively moral emotions and more on how twinges of affect ground more broadly normative, evaluative, or practical judgments about what one should do, whether it concerns the appropriateness of white lies or green apples. After all, Hume did argue that passion drives not only moral judgment but all action by providing evaluative content—e.g., that white lies are *acceptable* or that the Granny Smith apples are a *worse* option than the Opals.

Without proper function of the amygdala or vmPFC, one might argue, we could not even assign value to an object or event (Railton 2017; Carruthers 2018).

However, affect is not specific to judgments about what's best to do. All manner of cognition is suffused with affect, including attention, visual experience, memory encoding and retrieval, and semantic processing (Duncan & Barrett 2007; Hohwy 2013: ch. 12; Woodward 2016; Seligman et al. 2016; Carruthers 2018). Although we've focused on patients with deficits in decision-making, we could illustrate the ubiquity of affect by discussing patients with delusions or perceptual irregularities due to affective impairment in areas like the amygdala and vmPFC. Individuals with Capgras syndrome, for example, tend to lack the feeling of familiarity when seeing the faces of loved ones and thus claim that their family members have been replaced with imposters. These patients' brains appear to suffer from poor communication between the amygdala and face-processing areas in the temporal lobe (see e.g. Hirstein & Ramachandran 1997). In neurotypical individuals, by contrast, affect seems to provide significance and meaning to the visual experience of familiar faces, which facilitates the inference to another person's identity. In this way, affect doesn't specifically underwrite decision or action in the brains of humans (and other animals) but rather a broad notion of *significance*, from the meanings of words to the familiarity of faces or places. These are domains of cognition that sentimentalists have traditionally aimed to distinguish from morality, aesthetics, and prudence.

At any rate, problems arise for classical sentimentalism even if we concede that affect is emotional and indeed that it provides as input the distinctive value component to evaluative judgment. As sentimentalists themselves have been quick to point out (e.g. Nichols 2004), value alone is neither specific to nor sufficient for moral judgment. Humans and many other animals use their amygdala and frontal cortices to assign positive and negative values to many types of non-moral events, such as wildfires and levers that yield a food reward. Sustenance is often good and wildfires are often bad, but they aren't always morally wrong or the best thing to bring about. Casualties, property damage, and habitat loss are intrinsically bad and disvalued, but it might be better, all things considered, to avoid nuclear war than stop the wildfire. Similarly, attaining food might be intrinsically valued, and so rewarding, yet a rat will refrain from eating because it more greatly fears the punishment it has learned to associate with eating from a particular trough. What's missing—and what's essential to moral judgment—is something like the application of a rule or similar categorization, which doesn't seem to be captured by valuation alone.

Now, the defense of rationalism so far may seem weak if it turns out that moral cognition is in some respects special. Perhaps it is supported by a dedicated module that involves more affect than inference. In the next section, however, we'll see that converging evidence suggests that moral cognition is supported by a spatially distributed network of brain regions that engage in domain-general processes that are arguably inferential.

4. Step 2: Domain-General Reasoning

So far, we've seen that moral judgment is at least subserved by two brain regions often associated with emotion: the amygdala and vmPFC. However, an improved understanding of these and other areas suggests that the "emotion" here is largely unconscious twinges of affect which facilitate many cognitive functions, including learning and reasoning. The entanglement of inference and affect reveals the centrality of reasoning in moral judgment, even if it is often automatic and unconscious. The neurobiological support for moral rationalism, however, is not yet complete, for there is more evidence to consider. In this next step, we'll examine further research that reveals a more complete picture of the moral brain. What emerges is a set of regions that subserve general rational capacities, ones we'd antecedently expect to be deployed in ethical thought.

4.1 Inferring Intent

There are multiple methods one can use to study different aspects of moral judgment, and we'll encounter a number of different approaches. Let's begin with moral evaluations of accidental versus intentional harms. We often evaluate the morality of an action by whether it caused harm, but the adage "No harm, no foul" only applies in limited circumstances. As with attempted murder, malicious intent can be enough to make an act morally objectionable (indeed criminal), even if it ultimately caused no harm. Similarly, even when there is a terrible outcome, such as a heartbreaking loss of life, we judge the culprit less harshly if the harm was accidental, even if negligent. In this way, intent interacts with outcomes to influence many moral judgments.

Although assessing intent isn't the only core aspect of ethical reasoning (Graham et al. 2013; Young & Saxe 2011; Parkinson et al. 2011), it is, to some degree or other, commonplace in moral cognition across most societies (Barrett et al. 2016; McNamara et al. 2019). It's part of a general capacity for inferring the mental states of other individuals or "theory of mind" (Decety & Lamm 2007; Young & Dungan 2012), which is primarily subserved by two overlapping brain areas: the *temporoparietal junction* or *posterior* portions of the *superior temporal sulcus* (TPJ/pSTS). Naturally, researchers have hypothesized that this circuit is involved in reasoning about intent in moral judgment.

To test this, participants evaluated hypothetical, yet realistic, scenarios in which intent and outcome varied. The vignettes involve someone who harms another accidentally or intentionally or merely attempts to do so (see Table 1). For example, in one of the Attempted Harm cases, Grace and her friend are touring a chemical plant when Grace puts what she believes is a toxic substance in her friend's coffee, though in fact it's sugar. Would the temporoparietal junction be more active than other brain areas while participants evaluated such scenarios in the scanner? The answer, as predicted, is yes, and the TPJ is most active when evaluating cases of merely attempted harm, where there is no bad outcome and a negative moral judgment rests heavily on the agent's malicious intent (Young et al. 2007; Young & Saxe 2008).

Table 1: **Summary of Vignettes that Vary Intention and Outcome**
(Adapted from Young et al. 2007)

	Negative Outcome	*Neutral Outcome*
Negative Intention	**Intentional Harm**: Grace believes the substance is *poison* and it is *poison* (so her friend *dies*).	**Attempted Harm**: Grace believes the substance is *poison* but it is *sugar* (so her friend *lives*).
Neutral Intention	**Accidental Harm**: Grace believes the substance is *sugar* but it is *poison* (so her friend *dies*).	**Neutral**: Grace believes the substance is *sugar* and it is *sugar* (so her friend *lives*).

Of course, such neuroimaging studies only provide correlations between moral judgment and brain circuits. But further research provides reason to believe that the TPJ is causally implicated in such moral cognition. In particular, evaluations of merely attempted harms are less harsh when the TPJ is disrupted non-invasively through transcranial magnetic stimulation (Young et al. 2010). Moreover, moral judgments about such scenarios appear to be somewhat abnormal in individuals with autism, a condition with characteristic deficits in theory of mind (Moran et al. 2011; see also Koster-Hale et al. 2013).

Importantly, theory of mind is a domain-general capacity. It can be deployed in areas well beyond ethics, such as learning to dance with a partner or figuring out how to intimidate an opponent on the football field. So there is converging neurobiological evidence that the TPJ, with its domain-general mind-reading abilities, plays a key role in a core form of moral assessment. This, of course, is just one relevant domain-general capacity in moral cognition and, accordingly, only one key brain circuit.

4.2 Applying Rules, Calculating Consequences

Another paradigm in moral neuroscience examines how we trade off different values in the context of moral dilemmas. In these sacrificial dilemmas, one must indicate whether it is morally appropriate to promote the greater good (e.g., stopping a runaway trolley from harming multiple individuals) at the expense of inflicting harm on the few (e.g., stopping the trolley requires pushing a single large person in front of it). Although the famed trolley problem is a common vignette in this literature, it's important to recognize that researchers employ other more realistic sacrificial dilemmas as well, such as a scenario in which one can encourage the use of a vaccine that will harm some citizens but save many others (see the "Vaccine Policy" case in the supplemental materials of Greene et al. 2001). Here we see a classic conflict in ethics between utilitarian or consequentialist concerns about outcomes and non-consequentialist concerns about upholding general rules or principles, such as those that prohibit treating others as if they were mere objects to be used as a means to one's ends.

Joshua Greene and his collaborators suspected such moral conflicts arise from another domain-general system, or rather two systems, in the human brain. According to the familiar dual-process theory of cognition, humans have two rather different modes of thought. One generates fast, automatic, intuitive judgments while the other produces slow, reflective, deliberations—"thinking, fast and slow," as Kahneman (2011) puts it.

We can see this in action across all of human cognition, from language to mathematics and even social norms. Consider, for example, a famous mathematical problem on the Cognitive Reflection Test: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" (Frederick 2005). The intuitive answer is 10 cents, but some deliberation should yield 5 cents. The same tension in thought arises in other domains as well, such as social norms related to naming: "Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?" (Thomson & Oppenheimer 2016). (Test yourself: Which is the automatic intuitive response and which is the correct response you get upon reflection?) The phenomenon is not restricted to word problems either. Consider, finally, the well-known Stroop task in which one is instructed to name the ink color that a series of color-words are written in (e.g. "red" written in purple), rather than name the meaning of the word (Stroop 1935). Modern literate people are so habituated to reading words that they must deliberately override the automatic impulse when faced with this task, and thus they must slow down to avoid mistakes. Other linguistic capacities likewise become habitual or automated, including the ability to discern the grammaticality of sentences in one's native language (Mikhail 2011). These linguistic examples show that intuitive responses are not necessarily unlearned, innate intuitions but often based in large part on prior experience, habit, and cognitive automation, through both conscious and unconscious inference. Indeed, if one practices enough, the Stroop task and Cognitive Reflection Test become less arduous.

Specific neural circuits appear to be dedicated to these two systems of thought. Although a number of brain areas are involved, two of the key players reside next to each other in the prefrontal cortex. One of these, the vmPFC, which we've already encountered, is linked to automatic intuitive responses. The other, the *dorsolateral prefrontal cortex* (dlPFC), is associated with slow deliberation and the regulation of automatic thoughts (Miller & Cohen 2001; McClure et al. 2004). Importantly, these are domain-general functions that can be applied to judgments or decisions about grammar, social norms, math problems, and more.

The insight from Greene and his collaborators was that moral cognition likely exhibits the same dual-system characteristics, particularly when we evaluate moral dilemmas. Moreover, the researchers provide evidence that responses to such moral dilemmas invoke the usual cast of neural circuits in dual process theory (Greene et al. 2001; 2004). When participants gave automatic intuitive responses to a moral dilemma (don't sacrifice the few for the greater good), they exhibited elevated activity in the vmPFC, as well as the amygdala, pSTS/TPJ, and the *posterior cingulate cortex* (PCC). Heightened activation in the PCC is unsurprising given that it's associated with, among other functions, resolving conflicts. When participants provided the counter-intuitive response to a moral dilemma (the needs of the many outweigh the needs of the few), they exhibited greater activity in the dlPFC, as well as the pSTS/TPJ/inferior parietal lobe (but compare Kahane et al. 2012).

Now, Greene (2014) argues further, and more controversially, that these two brain areas and modes of cognition are tied to specific forms of moral values (characteristically utilitarian vs. non-utilitarian moral judgments). However, for our purposes, it is only important to show that core forms of moral judgment recruit the sorts of brain areas implicated in two domain-general modes of cognition—moral thinking, fast and slow. In the context of moral judgment, the dlPFC may be more involved in calculating the outcomes of an action while the vmPFC may help apply general moral rules (or assigning values to act types). But we needn't construe these processes as characteristic of certain ethical theories.

Greene (2014) does marshal a wide range of neurobiological evidence to support this dual-process architecture of moral cognition in the brain. Importantly, some of the evidence goes beyond mere correlations found in brain imaging experiments, by drawing on lesion studies and the manipulation of slow, deliberate moral thinking. Consider, for example, studies of the moral judgments of people with deficits in their automatic emotional responses to ethically charged situations, such as individuals with psychopathy, damage to the vmPFC, and frontotemporal dementia (Koenigs et al. 2012; Koenigs et al. 2007; Mendez et al. 2005). As the dual-process model predicts, these patients tend to give more counter-intuitive responses to the sacrificial dilemmas (although it's unclear whether this is due to a "utilitarian" concern for the greater good or anti-social tendencies to accept instrumental harm; see Kahane et al. 2018).

Greene also asserts, more controversially, that automatic moral intuitions are often inflexible and unreliable, at least when applied to novel moral problems to which the intuitions were not attuned. He argues, for example, that our automatic responses to sacrificial moral dilemmas inflexibly respond to how promoting the greater good requires pushing and other forms of "personal force" or "prototypical violence" (contrast Feltz & May 2017).

However, ample evidence suggests that our automatic moral intuitions—like automatic responses in the Stroop task and our intuitive judgments of grammaticality—are frequently the product of sophisticated unconscious learning and reasoning over time (Mikhail 2011; Railton 2017; May 2018a; Nichols 2021). Indeed, our automatic intuitions, both within and outside of the domain of ethics, are not fixed but rather malleable responses to changes in one's environment and circumstances. As we've seen, brain areas like the vmPFC and the amygdala used to be associated exclusively with basic emotional responses (e.g., fear), which is why in early work some dual-process theorists identified automatic intuitions with emotions and controlled deliberation with reason (e.g. Greene et al. 2001; contrast Greene 2014). But it is now clear (see §3) that these circuits serve more general learning and inferential capacities. Affect plays a crucial role in these processes of learning and inference, but again such twinges of positive or negative thought, which are not always felt, are best construed as signals of significance, not emotions—let alone paradigm moral emotions, such as compassion or indignation. Thus, the least controversial aspects of the dual-process model suggest that when resolving dilemmas moral cognition, whether fast and slow, involves domain-general capacities.

4.3 Moral vs. Non-Moral Statements

The forgoing research in the neuroscience of ethics has received much attention and yet curiously focuses on rather specific kinds of moral judgment. Both approaches study only harm, leaving out other moral foundations, such as loyalty and fairness (see Graham et al. 2013; Parkinson et al. 2011). It is remarkable, nevertheless, that we already see a similar cast of neural circuits cropping up in the two distinct lines of research. Moreover, another research paradigm provides even further converging evidence using a broader range of moral judgments.

This series of neuroimaging studies attempted to identify which brain regions are more active when people assess statements about a range of moral phenomena, not sacrificial dilemmas and not exclusively situations involving harm (Moll et al. 2001; Moll et al. 2002). While in the brain scanner, participants judged as right or wrong:

- moral statements (e.g., "The father never treated his son as a slave," "The elderly are useless"),
- non-moral statements (e.g., "The painter used his hand as a paintbrush," "The elderly sleep more at night"), and
- scrambled statements (e.g., "Sons push use eat work.").

Remarkably, the researchers found that the moral judgments primarily produced greater activity in our familiar set of neural circuits: portions of the prefrontal cortex, amygdala, temporoparietal junction, and cingulate. It's striking that such different approaches and methods in moral neuroscience have zeroed in on roughly the same set of brain areas.

5. The Defense of Rationalism

Reviews of multiple studies and methods suggest that we are beginning to uncover a well-corroborated picture of the moral brain (Moll et al. 2005; Greene 2009; Demaree-Cotton & Kahane 2018; May et al. 2022). The circuits form a spatially distributed network that, at least provisionally, subserve some core forms of moral judgment (see Figure 1). But talk of "the moral brain" is merely a stylistic device. Rather than a distinct moral module, we see domain-general circuits that facilitate moral and prudential cognition (Arvan 2020) and much more still.

Importantly for rationalism, the relevant areas are associated with a number of rational capacities, particularly:

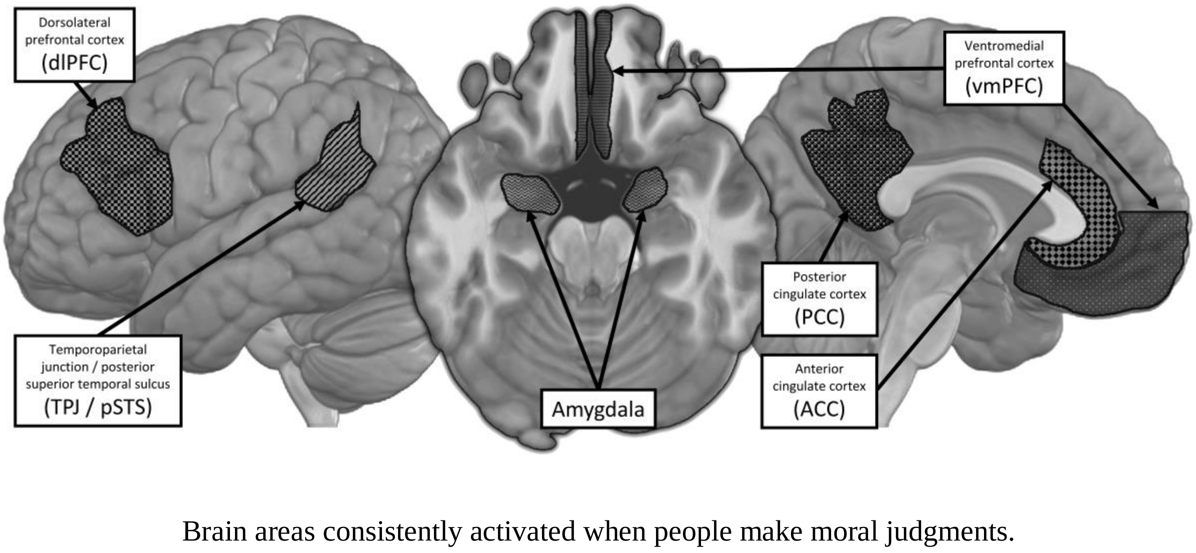
- Assigning learned values or rules to act types (**amygdala**)

- Calculating the consequences of an action and integrating that information with general rules or heuristics (portions of the **prefrontal cortex**)
- Computing the level of intent behind an action (**temporoparietal junction**).
- Detecting conflicts and errors in one's expectations (portions of the **cingulate**)

These are not just any set of domain-general capacities but ones we would antecedently expect to be applicable, even central, to moral cognition. Of course, these capacities deployed in moral judgment are rather automatic and unconscious. Do they genuinely facilitate reasoning or inference, rather than mere computation? Unfortunately, the scientific evidence doesn't settle that thorny question. For our purposes, it is enough that these domain-general capacities aren't associated with paradigm moral emotions and could very well involve or facilitate inference. It might seem implausible that rather automatic and unconscious processes facilitate inferential transitions among beliefs or belief-like states, but that is an increasingly prominent view among a range of philosophers, neuroscientists, and other cognitive scientists (see e.g. Hohwy 2013; Mandelbaum 2016; Buckner 2019).

How do these circuits work together? Further research is needed, but some researchers have gained an understanding of the order in which such brain areas lead to moral judgment (at least of harmful acts), by using EEG. While participants wore an electrode-filled cap and viewed short videos of intentional versus accidental harms, researchers found that in a mere 60 milliseconds participants inferred the agent's mental states using the TPJ/pSTS (Decety & Cacioppo 2012). In another EEG study concerning both harmful and helpful acts, in just a few hundred milliseconds after activity in the TPJ/pSTS, the amygdala appears to provide evaluative input to areas of the prefrontal cortex, at which point a moral judgment is made (Yoder & Decety 2014). As neuroimaging researchers have likewise found (Shenhav & Greene 2014), the amygdala seems to provide an assessment of the positive or negative value of the act in question, which the prefrontal cortex then integrates with information about harmful outcomes in order to make an overall moral judgment.

Figure 1: Key Moral Circuits


Dorsolateral prefrontal cortex (dlPFC)
Temporoparietal junction / posterior superior temporal sulcus (TPJ / pSTS)
Amygdala
Ventromedial prefrontal cortex (vmPFC)
Posterior cingulate cortex (PCC)
Anterior cingulate cortex (ACC)

Brain areas consistently activated when people make moral judgments.
(Source: May et al. 2022)

Compare this model of moral judgment with other more domain-*specific* cognitive functions. Sometimes damage to a neural circuit seems to impair rather specific capacities, such as the ability to identify animate but not inanimate objects. For instance, just months after traumatic brain injury or stroke, some patients struggle to name fruits and vegetables but not animals (see patient P.S. in Caramazza & Shelton 1998: 3). We don't tend to find such domain-specific deficits in moral cognition. Damage to the vmPFC might seem to selectively disrupt the normal ability to condemn instrumental harm—e.g., judging it inappropriate to kill one person as a means to saving five others (Koenigs et al. 2007). But a more complete picture of the evidence suggests that dysfunction in the vmPFC disrupts a more general-purpose learning mechanism and accordingly exhibits a wider range of deficits. Psychopaths and patients with "acquired sociopathy," for example, don't just provide abnormal moral judgments—and certainly not just a propensity to disvalue instrumental harm—but also deficits in prudence, recognizing others' mental states, learning from punishment, and even deciding which kind of apples to buy at the grocery store.

One upshot of this analysis is that it can be misleading to describe the dispute between sentimentalists and rationalists in terms of analogies with aesthetics and mathematics. These analogies are initially helpful in illustrating the two camps (Gill 2007). But we shouldn't saddle rationalism with predicting, for example, that moral and aesthetic judgments won't recruit similar brain areas (cf. Heinzelmann et al. 2020). The emerging neuroscience, as well as philosophical analysis of the dispute, suggests that rationalists needn't make such predictions, given that affect facilitates inference (entanglement) and does so across the board, from mathematics and language to ethics and aesthetics (domain-generality).

6. Conclusion

Psychological versions of sentimentalism and rationalism do make empirical predictions and are thus partly accountable to scientific findings. I have argued that distinct lines of research in neuroscience are converging on a picture of moral circuitry in the brain that coheres well with moral rationalism. Novel forms of sentimentalism might also be made consistent with the data. Indeed, there is a sense in which our understanding of the human brain makes it difficult to distinguish modern forms of the two theories. Ultimately, rationalism might fare better than sentimentalism if we wish to sharply distinguish the two camps, though my primary aim has only been to show that a traditional form of rationalism remains defensible.

The argument involved two main steps. The first step drew on studies of brain abnormalities and psychopathologies to show that inference and affect are *entangled* in moral cognition (as well as other domains). It may be tempting for a sentimentalist to point to the centrality of the amygdala and vmPFC and declare victory, given that these regions were traditionally associated with emotion. But we're learning that these areas, although suffused with affect, subserve mechanisms for learning and reasoning. The second step of the argument focused on how circuits in the human brain contribute *domain-general* processes to neurotypical moral evaluation. A wide range of studies reveal a suite of spatially distributed neural circuits known to facilitate domain-general capacities, such as mind-reading and calculating consequences.

Importantly, the neurobiological argument presented here draws on a broad range of evidence. We have not focused on a small set of provocative studies, one methodology, or work from a few labs using one paradigm. We've encountered evidence from neuroimaging, psychopathology, and even neurostimulation to provide evidence of causal relationships. Nevertheless, the model of the moral brain presented here is certainly provisional and incomplete. As I've stressed, most of the studies focus on harm and fairness, leaving other moral values understudied. Equally important to recognize, however, is how striking it is that roughly the same cast of circuits appears throughout diverse approaches to moral neuroscience. The cumulative evidence, I've argued, fits well with the rationalist tradition in moral psychology, which treats morality like other domains of thought in its deployment of domain-general reasoning capacities.

Acknowledgements

Versions of this article were presented at the American Philosophical Association (Central Division), International Neuroethics Society, Italian Society for Neuroethics (receiving the Best Paper Prize in Philosophy), Mississippi State University, the Department of Educational Studies in Psychology at the University of Alabama, and neuroimaging lab meetings at the University of Alabama (Rajesh Kana) and the University of Alabama at Birmingham (Kristina Visscher). Thanks to the organizers and participants for their incisive feedback and to Mara Bollard, Joseph McCaffrey, and Mohammad Waqas for comments on the manuscript. This project was made possible through the support of a grant from the John Templeton Foundation (Academic Cross-Training Fellowship 61581). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

References

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2012). Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology*, 121(2), 484–497.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC.

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience* 2(11), 1032–1037.

Arvan, M. (2020). *Neurofunctional Prudence and Morality: A Philosophical Theory*. Routledge.

Barrett, H. C., Bolyanatz, A., Crittendend, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17), 4688–4693.

Berker, S. (2009). The Normative Insignificance of Neuroscience. *Philosophy & Public Affairs*, 37(4), 293–329.

Betancourt, H. 1990. "An Attribution-Empathy Model of Helping Behavior." *Personality and Social Psychology Bulletin* 16(3): 573–591.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. Trends in Cognitive Sciences, 11(9), 387–392.

Boghossian, P. (2012). "What is Inference?" *Philosophical Studies* 169(1): 1–18.

Buckner, C. (2019). Rational inference: The lowest bounds. *Philosophy and Phenomenological Research* 98(3): 697-724.

Caramazza, A. & Shelton, J. R. (1998). Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience* 10, 1–34.

Carruthers, P. (2018). Valence and value. *Philosophy and Phenomenological Research* 97(3): 658-680.

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.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

D'Arms, J. & Jacobson, D. (2014). Sentimentalism and Scientism. In J. D'Arms & D. Jacobson (Eds.), *Moral Psychology and Human Agency*. Oxford University Press.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of Neurophysiology*, 108(11), 3068–3072.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13(6), 580–593.

Demaree-Cotton, J., & Kahane, G. (2018). The Neuroscience of Moral Judgment. In A. Z. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge Handbook of Moral Epistemology*. Routledge.

Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition and emotion* 21(6): 1184-1211.

Eagleman, D. (2015). *The Brain: The Story of You*. Pantheon.

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition*, 166, 314–327.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives* 19(4): 25-42.

Gill, M. B. (2007). Moral Rationalism vs. Moral Sentimentalism: Is Morality More Like Math or Beauty? *Philosophy Compass*, 2(1), 16–30.

Glenn, A. L., & Raine, A. (2014). *Psychopathy*. New York University Press.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. 2013. "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism." In *Advances in Experimental Social Psychology* 47: 55–130.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press, pp. 987-999.

Greene, J. D. (2014). Beyond Point-and-Shoot Morality. *Ethics*, 124(4), 695–726.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, 44(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108.

Haas, J. (2020). Moral Gridworlds: A Theoretical Proposal for Modeling Artificial Moral Cognition. *Minds and Machines* 30: 219–246.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail. *Psychological Review*, 108(4), 814–834.

Hare, R. D. (1993). *Without Conscience: The Disturbing World of the Psychopaths Among Us*. Guilford Press.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience* 28(22): 5623-5630.

Heinzelmann, N. C., Weber, S. C. & Tobler, P. N. (2020). Aesthetics and morality judgments share cortical neuroarchitecture. *Cortex* 123: 484-495.

Hirstein, W., & Ramachandran, V. S. (1997). Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 264(1380): 437-444.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Horgan, T., & Timmons, M. (2007). "Morphological Rationalism and the Psychology of Moral Judgment." *Ethical Theory and Moral Practice* 10(3): 279–95.

Huebner, B. (2015). Do Emotions Play a Constitutive Role in Moral Cognition? *Topoi*, 34(2), 427–440.

Hume, D. (1739–40/2000). *A Treatise of Human Nature*, ed. by D. F. Norton & M. J. Norton. Oxford University Press.

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience*, 35(36), 12593-12605.

Jones, K. (2006). "Quick and Smart? Modularity and the Pro-Emotion Consensus." *Canadian Journal of Philosophy* 36(Supplement): 3–27.

Jones, K., & Schroeter, F. eds. (2018). *The Many Moral Rationalisms*. Oxford University Press.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social cognitive and affective neuroscience* 7(4): 393-402.

Kahane, G., Everett, J. A., 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.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kauppinen, Antti (2018). Moral Sentimentalism. *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2018/entries/moral-sentimentalism

Kauppinen, A. (2019). What Is Sentimentalism? What Is Rationalism? *Behavioral and Brain Sciences* 42(e146):25-27.

Kennett, J. (2006). Do psychopaths really threaten moral rationalism? *Philosophical Explorations*, 9(1), 69–82.

Kennett, J., & Fine, C. (2008). Internalism and the Evidence from Psychopaths and "Acquired Sociopaths." In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3*. MIT Press, pp. 173–190.

Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research*, 142(2-3), 107–128.

Kiehl, K. & Sinnott-Armstrong, W. eds. (2013). *Handbook of Psychopathy and Law*. Oxford University Press.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, 7(6), 708–714.

Koenigs, M., Young, L. L., Adolphs, R., Tranel, D., Cushman, F. A., Hauser, M. D., & Damasio, A. R. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements, *Nature* 446(7138), 908–911.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding Moral Judgments From Neural Representations of Intentions. *Proceedings of the National Academy of Sciences*, 110(14), 5648–5653.

Kumar, V. & May, J. (2019). How to Debunk Moral Beliefs. In Methodology & Moral Philosophy, J. Suikkanen & A. Kauppinen (eds.). Routledge, pp. 25–48.

Machery, E., & Doris, J. M. (2017). "An Open Letter to Our Students: Doing Interdisciplinary Moral Psychology." In *Moral Psychology: A Multidisciplinary Guide*, ed. by B. G. Voyer and T. Tarantola, pp. 127–47. Springer.

Maibom, H. L. (2005). Moral Unreason: The Case of Psychopathy. *Mind & Language* 20(2), 237–257.

Maibom, H. (2010). "What Experimental Evidence Shows Us About the Role of Emotions in Moral Judgement." *Philosophy Compass* 5 (11):999-1012.

Malle, B. F., & Scheutz, M. (2019). Learning How to Behave: Moral Competence for Social Robots. In: Bendel O. (eds) *Handbuch Maschinenethik*. Springer VS, Wiesbaden, pp. 255-278.

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs* 50(3): 629-658.

May, J. (2018a). *Regard for Reason in the Moral Mind*. Oxford University Press.

May, J. (2018b). "The Limits of Appealing to Disgust." *The Moral Psychology of Disgust*, ed. by V. Kumar & N. Strohminger. Rowman & Littlefield.

May, J. (2021). Bias in Science: Natural and Social. *Synthese*.

May, J. & Kumar, V. (2018). "Moral Reasoning and Emotion." *The Routledge Handbook of Moral Epistemology*, eds. K. Jones, M. Timmons, & A. Zimmerman, Routledge.

May, J., Workman, C., Haas, J. & Han, H. (2022). The Neuroscience of Moral Judgment: Empirical and Philosophical Developments. *Neuroscience & Philosophy*, eds. F. de Brigard & W. Sinnott-Armstrong. MIT Press.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science* 306(5695): 503-507.

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing Outcome vs. Intent Across Societies: How Cultural Models of Mind Shape Moral Reasoning. *Cognition*, 182, 95–108.

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An Investigation of Moral Judgement in Frontotemporal Dementia. *Cognitive and Behavioral Neurology*, 18(4), 193–197.

Mikhail, J. (2011). *Elements of Moral Cognition*. Cambridge University Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience* 24(1): 167-202.

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

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional Networks in Emotional Moral and Nonmoral Social Judgments, *NeuroImage* 16(3), 696–703.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The Neural Basis of Human Moral Cognition. *Nature Reviews Neuroscience* 6(10), 799-809.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired Theory of Mind for Moral Judgment in High-Functioning Autism. *PNAS*, 108(7), 2688–2692.

Nichols, S. (2004). *Sentimental Rules*. Oxford University Press.

Nichols, S. (2021). *Rational Rules: Towards a Theory of Moral Learning*. Oxford University Press.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is Morality Unified? Evidence That Distinct Neural Systems Underlie Moral Judgments of Harm, Dishonesty, and Disgust. *Journal of Cognitive Neuroscience*, 23(10), 3162–3180.

Plakias, A. (2018). The Response Model of Moral Disgust. *Synthese* 195 (12):5453-5472.

Prinz, J. J. (2006). The Emotional Basis of Moral Judgments. *Philosophical Explorations* 9(1): 29-43.

Prinz, J. J. (2016). Sentimentalism and the Moral Brain. S. M. Liao (ed.) *Moral Brains*. Oxford University Press, pp. 45-73.

Railton, P. (2017). Moral Learning: Conceptual Foundations and Normative Relevance. *Cognition*, 167(Oct), 172–190.

Rozin, P., Markwith, M., & Stoess, C. (1997). "Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust." *Psychological Science* 8(2): 67-73.

Salzman, C. D., & Fusi, S. (2010). Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annual review of neuroscience* 33: 173-202.

Sauer, H. (2017). *Moral Judgments as Educated Intuitions*. MIT Press.

Saver, J. L., & Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, 29(12), 1241–1249.

Siegel, S. (2017). *The rationality of perception*. Oxford University Press.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. S. (2016). *Homo Prospectus*. New York: Oxford University Press.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, 34(13), 4741-4749.

Sinhababu, Neil (2017). *Humean Nature: How Desire Explains Action, Thought, and Feeling*. Oxford University Press.

Stroop, J. Ridley (1935). Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology* 18 (6): 643.

Taber-Thomas, B. C., Asp, E. W., Koenigs, M., Sutterer, M., Anderson, S. W., & Tranel, D. (2014). Arrested development: early prefrontal lesions impair the maturation of moral judgement. *Brain*, 137(4), 1254–1261.

Tappolet, C. (2016). *Emotions, Value, and Agency*. Oxford University Press.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an Alternate Form of the Cognitive Reflection Test. *Judgment and Decision Making* 11(1): 99-113.

Woodward, J. (2016). Emotion versus Cognition in Moral Decision-Making. In *Moral Brains: The Neuroscience of Ethics*, ed. by S. Matthew Liao. Oxford University Press, pp. 87–117.

Yang, Q., Yan, L., Luo, J., Li, A., Zhang, Y., Tian, X., and D. Zhang. 2013. "Temporal Dynamics of Disgust and Morality: An Event-Related Potential Study." *PloS one* 8(5): e65094.

Yang, Q., Li, A., Xiao, X., Zhang, Y., & Tian, X. 2014. "Dissociation between morality and disgust: An event-related potential study." *International Journal of Psychophysiology* 94(1):84-91.

Yoder, K. J., & Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. *Neuropsychologia*, 60, 39–45.

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

Young, L. L., Cushman, F. A., Hauser, M. D., & Saxe, R. (2007). The Neural Basis of the Interaction between Theory of Mind and Moral Judgment. *Proceedings of the National Academy of Sciences*, 104(20), 8235–8240.

Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social neuroscience*, 7(1), 1-10.

Young, L. L., & Saxe, R. (2008). The Neural Basis of Belief Encoding and Integration in Moral Judgment. *NeuroImage* 40(4), 1912–1920.

Young, L. L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, 120(2), 202–214.