The Neuroscience of Moral Judgment: Empirical and Philosophical Developments

Joshua May, Clifford I. Workman, Julia Haas, & Hyemin Han

Forthcoming in *Neuroscience and Philosophy*,
eds. Felipe de Brigard & Walter Sinnott-Armstrong (MIT Press).

Abstract: We chart how neuroscience and philosophy have together advanced our understanding of moral judgment with implications for when it goes well or poorly. The field initially focused on brain areas associated with reason versus emotion in the moral evaluations of sacrificial dilemmas. But new threads of research have studied a wider range of moral evaluations and how they relate to models of brain development and learning. By weaving these threads together, we are developing a better understanding of the neurobiology of moral judgment in adulthood and to some extent in childhood and adolescence. Combined with rigorous evidence from psychology and careful philosophical analysis, neuroscientific evidence can even help shed light on the extent of moral knowledge and on ways to promote healthy moral development.

Key Words: moral cognition, moral learning, moral development, moral epistemology, trolley problem

1. Introduction

Imagine reading in the news about a country far, far away that won't admit poor, helpless, asylum-seeking refugees. "That's just *wrong*," you think. Yet your neighbor passionately draws the opposite conclusion. In this way, humans are able to judge actions to be right or wrong and people to be good or evil, even if those criticized (or praised) are "third parties" with no connection to the individual making the judgment. Another interesting feature of moral judgment is that, even if it usually motivates one to follow it, the corresponding action often fails to materialize. Many people lie, cheat, and steal yet feel guilty afterward because they believe such acts are immoral. So, while human moral psychology has many components—including motivation, decision, and action—our focus is on *moral judgment* specifically.

How does this capacity arise in the brain? In this chapter, we identify and weave together some of the major threads of research emerging over the past 30 years, before discussing future directions. We'll explain how the first thread in the neuroscience of moral judgment focused on brain abnormalities, from lesion studies to disorders such as psychopathy. Early research (ca. 1990-2000) concentrated on the amygdala and ventromedial prefrontal cortex, which led to theories emphasizing gut feelings as integral to moral judgment. The second thread

(ca. 2000-2010) ushered in new methods, especially brain imaging, brain stimulation, and neurotransmitter manipulations. Moral neuroscience highlighted brain areas associated with complex computation and reasoning, such as the dorsolateral prefrontal cortex (dlPFC) and the temporal parietal junction(TPJ)/posterior superior temporal sulcus (pSTS). Theorists introduced dual process models to explain how both gut feelings and conscious deliberation influence moral cognition. More recent trends have drawn more on animal models and computational mechanisms to develop theories of moral learning in terms of reward and valuation. We expect the future will include more work on how brain development from childhood to adolescence relates to moral cognition.

Ultimately we'll see that neuroscience, when combined with findings from psychology and allied disciplines, helps to address perennial philosophical questions about moral judgment, particularly the possibility and limits of knowing right from wrong. By helping to uncover the neural circuits and neurocognitive mechanisms underlying mature moral cognition, how it normally develops, and how it breaks down in pathology, neuroscience sheds light on the trustworthiness of our moral judgments and the possibility and shape of moral progress. It's by no means easy to bridge the dreaded gap between how we *do* form our moral beliefs and how we *ought* to. Nevertheless, with caution and critical reflection, neuroscience can enrich our understanding of moral judgment and when it works well or poorly.

2. First Thread: Gut Feelings

In the 19th and 20th centuries, scientists and physicians inferred brain function primarily by examining patients with brain abnormalities, whether due to freak accidents, genetic anomalies, neurosurgery to treat debilitating seizures, or diseases like herpes that can damage the nervous system. When lesions disrupted patients' moral capacities, researchers inferred that the affected brain area facilitates the corresponding capacity.

1.1 Somatic Markers

The most famous, even if poorly documented, case is that of Phineas Gage. In 1848, while working on a railroad in Vermont, an explosive accident catapulted a 3-foot iron rod up through Gage's left cheek and out the top of his head. Remarkably, he survived, but his personality reportedly changed so much that he was virtually unrecognizable as his former self. Unfortunately, there is little reliable, corroborating evidence about the details of Gage's case (Macmillan 2000), but some of the changes were ostensibly apparent in his moral character. He reportedly became more impulsive, vulgar, rash, and even childish.

Although the affected brain area is difficult to pinpoint precisely, Gage seemed to suffer significant damage to the *ventromedial prefrontal cortex* (vmPFC), which overlaps with the orbitofrontal cortex (OFC) behind the eyes ("orbits"). We know much more now about individuals with damage to this area in adulthood. Patients allegedly develop what Antonio Damasio (1994) dubbed "acquired sociopathy." The label is misleading, though, because the psychological

profile of patients with vmPFC lesions is hardly similar to that of psychopaths (or what are sometimes referred to colloquially as "sociopaths"). Unlike individuals considered psychopathic, adults with vmPFC lesions are not callous, antisocial, or remorseless, but instead demonstrate a shortage of gut feelings that help guide decisions about what to do in the moment. Physiological measures suggest these patients, relative to controls, show diminished emotional responses when making a wide range of decisions, not just about how to treat others but even how to get points in a card game or where to eat for dinner. Patients with vmPFC damage generally give normal responses to questions about how various hypothetical choices should be resolved, including moral dilemmas (Saver & Damasio 1991), but they struggle to make decisions about what to do *oneself* in a *particular situation* (Kennett & Fine 2008). A patient might recognize it's impolite to talk about gory injuries at dinner, but does that mean the hilarious hiking story is off the table? Damasio attributes this deficit in decision-making to an underlying impairment of "somatic markers" that guide everyday decisions about how to behave. Without the relevant gut feelings, patients are able "to know but not to feel," as Damasio puts it (1994: 45).

1.2 Psychopathy

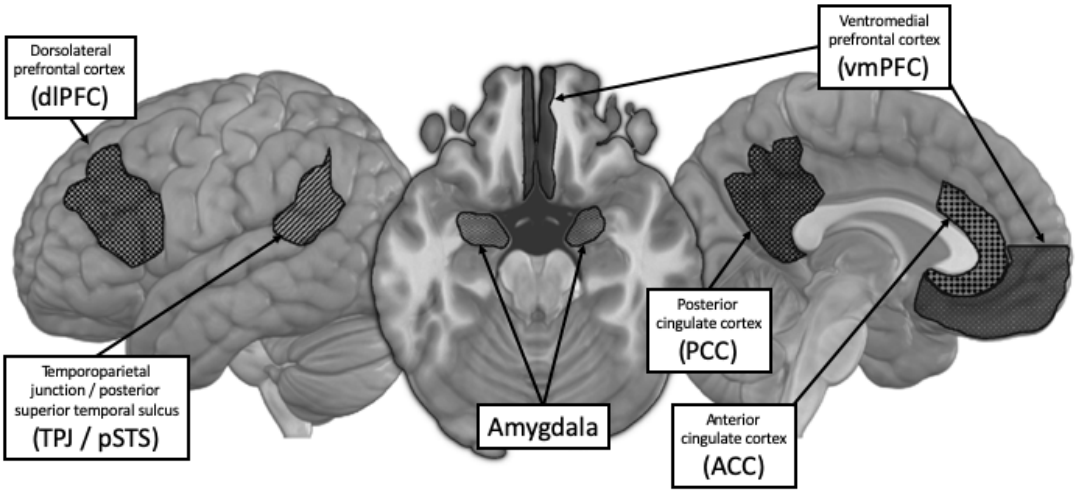
A more deviant example of abnormal moral thought and behavior lies in psychopathy. Psychopaths are characteristically callous, remorseless, manipulative, pompous, and exhibit superficial charm, among other similar vices that often leave injured or indigent victims in their wake (Hare 1993). Unlike "acquired sociopathy," individuals with psychopathy exhibit abnormal functioning in the vmPFC and amygdala (and their connectivity), along with other paralimbic areas associated with gut feelings (Blair 2007; Kiehl 2006). Part of the limbic system, the *amygdala* is a pair of small almond-shaped nodes deep in the brain (see Figure 1) and contribute, among other things, to assessing the significance of a stimulus, such as whether it is threatening, which plays a crucial role in learning. Unlike adults who acquire damage to vmPFC or amygdala (Anderson et al. 1999; Taber-Thomas et al. 2014), psychopathic traits are associated with abnormal structure and functioning in these regions (Kiehl & Sinnott-Armstrong 2013; Glenn & Raine 2014), whether due to unfortunate genes (e.g. alleles known to disrupt neurotransmitters) or adverse circumstances (e.g. childhood trauma, neglect, and even lead exposure).

Importantly, individuals with psychopathy do not only behave immorally; their understanding of right and wrong seems impaired to some extent. Some, but not all, incarcerated psychopaths exhibit some difficulty distinguishing moral rules from mere conventions (Aharoni et al. 2012). Interviews also suggest that some inmates with psychopathy have an inconsistent and tenuous grasp of moral concepts and reasons, particularly when attempting to justify decisions or to use relevant emotion words such as "guilt" (Hare 1993; Kennett & Fine 2008). Yet some people with psychopathic traits seem rational—indeed all too cold and calculating in their apparently skilled manipulation of others—although they do exhibit irrationality too, such as delusions of grandeur, poor attention span, and difficulty learning from punishment (Maibom 2005; May 2018). Thus, not only do the vmPFC and

amygdala seem to be crucial moral circuits, perhaps emotions are necessary for moral competence (Prinz 2016).

Figure 1: The Moral Brain

Brain areas consistently activated when people make moral, compared to non-moral, judgments.



1.3 Post-Hoc Rationalization

Also in the 1990s, moral psychologists began emphasizing gut feelings in moral judgment. Imagine being asked whether, hypothetically, it's morally acceptable for someone to clean a toilet with the national flag, to eat a pet that had been run over by a car, or to engage in consensual protected intercourse with an adult sibling. Most people in studies automatically condemn such "harmless taboo violations" without being able to articulate appropriate justifications (Haidt et al. 1993; Stanley et al. 2019). Indeed, it seems that we often intuitively regard actions as right or wrong first and only afterward does conscious reasoning concoct a defense of it (Haidt 2001; Cushman et al. 2006).

Similar ideas followed studies of split-brain patients, starting around the 1960s. When the corpus callosum is severed, often to treat seizures from epilepsy, the two hemispheres of the brain can no longer communicate with one another. Studies of such split-brain patients suggest that, in the absence of crucial information from one side of the brain, patients often confabulate a story to make sense of their behavior (Gazzaniga 1983). One commissurotomy patient, for instance, was tasked with choosing out of a row of eight images which two best relates to the two pictures recently presented to him. The catch was that only one picture was presented to each eye, and thus each side of the brain could process only one of the two pictures first presented. One half of the patient's brain saw a *home covered in snow* while the other half saw a *chicken claw*. To go with these, the patient almost instinctively chose an image of a *snow shovel* and an image of a *chicken head*. However, language abilities appear to be partly lateralized to one side of the brain, so one hemisphere can't readily communicate linguistically what it saw. As a result, the patient articulated a reason that appeared to be concocted just

to make sense of his intuitive choice, saying "you have to clean out the chicken shed with a shovel" (534). Recent work suggests callosotomy patients also provide confabulations in the context of moral judgments (Miller et al. 2010).

1.4 Lessons

In light of the above, and other studies in psychology and neuroscience, many theorists in the first thread adopted what we might call "sentimentalist" theories of moral judgment. Proponents asserted that moral attitudes and decisions are generated predominantly by automatic gut feelings, whereas reasoning is largely *post-hoc* rationalization (Haidt 2001; Nichols 2004; Prinz 2016). Now, some of these theorists took the evidence to erode the supposed division between reason and emotion (e.g. Damasio 1994) but with little emphasis on reasoning, inference, or complex computation underlying gut feelings and moral intuitions.

Sentimentalist theories do reconcile various observations of lesion patients, but several limitations remain. First, the centrality of gut feelings is *insufficiently corroborated*. While psychological studies initially appeared to support the importance of gut feelings in moral judgment, much of the findings were overblown (Huebner 2015; May 2018). One meta-analysis, for example, found limited evidence for the famous effect of incidental disgust priming on moral judgment, which disappeared entirely after controlling for publication bias (Landy & Goodwin 2015). Moreover, the vmPFC is unlikely a source of gut feelings but rather a hub wherein such feelings are incorporated with or weighed against other considerations before making a decision (Shenhav & Greene 2014; Hutcherson et al. 2015). Second, theories focusing on the vmPFC and amygdala are *incomplete*. Brain damage to rather different areas in the frontal and temporal lobes lead to moral dysfunction as well—e.g. in frontotemporal dementia. Moreover, as we'll see, subsequent brain imaging studies confirm that additional brain areas are integral to moral judgment. Indeed, early research on brain abnormalities often study the patient's social behavior and choice, rather than the moral evaluation of other people and their actions. Patients with vmPFC lesions, for example, don't appear to have distinctively moral deficits but rather problems with decision-making generally.

3. Second Thread: Reasoning

The second thread in moral neuroscience followed the development of neuroimaging technologies, which enabled the live, non-invasive measurement of brain functioning. Of these technologies, functional magnetic resonance imaging (fMRI) has dominated the methodological landscape. Moreover, partly given worries about post-hoc rationalization, researchers primarily investigated the neural correlates of moral judgments made in response to particular moral statements (Moll et al. 2001) or hypothetical scenarios (Greene et al. 2001). By varying the features of hypothetical scenarios, for instance, one can infer which factors shape moral judgments, instead of relying on the factors participants articulate as reasons for their moral judgments, which may be misleading.

3.1 Dual Process

The second thread arguably began with an influential article published by Joshua Greene and his collaborators (2001), in which participants underwent fMRI scanning while responding to *sacrificial dilemmas* familiar from longstanding philosophical debates in ethical theory. These hypothetical scenarios pit moral values against each other by describing an opportunity to sacrifice one person in order to save even more lives (typically five). Interestingly, most people consider it morally acceptable to sacrifice one to save five when this can be done in an *impersonal* way, such as diverting a runaway trolley away from five workers stuck on one track but onto another track with only one stuck worker (Switch scenario). But sacrificing one to save five is deemed unacceptable if the harm must be up close and *personal*, such as pushing someone in front of a trolley to save the five others (Footbridge scenario).

Such sacrificial dilemmas have long been used by philosophers to distinguish between and evaluate ethical theories. Treating personal harm as immoral, even when it would save more lives, ostensibly reflects characteristically "deontological" judgments that align with moral rules (e.g. don't kill), even if violating them would produce better consequences. Sacrificing one to save five, on the other hand, ostensibly reflects characteristically "utilitarian" (or consequentialist) judgments that privilege the maximization of overall welfare.

Greene's (2014) model adopts the tools of dual-process theory , which posits the operation of competing psychological processes, particularly automatic versus deliberative thinking (e.g. Kahneman 2011). Applying this to the moral domain, Greene theorizes that "utilitarian" responses to moral dilemmas are driven by controlled, deliberative reasoning while non-utilitarian ("deontological") responses are driven by automatic, intuitive, emotional heuristics that are relatively insensitive to the consequences of an action. Some of the support for this dual-process model comes from psychological experiments but also fMRI. Early on, Greene and his collaborators (2001; 2004) reported the engagement of predictably different brain areas are engaged when participants respond to personal and impersonal moral dilemmas. Compared to impersonal (and non-moral) dilemmas, personal dilemmas elicited greater activity in some areas associated with automatic, emotional, and social processing—namely, the vmPFC, amygdala, pSTS, and posterior cingulate cortex (PCC). Responses to impersonal dilemmas yielded greater activity in areas associated with controlled deliberative reasoning—namely, the dlPFC and inferior parietal lobe.

A diverse body of evidence appears to corroborate the dual-process model (Greene 2014), but let's focus on some of the brain science that goes beyond neuroimaging. Prior lesion studies appear consistent with the model's account of intuitive moral judgments being driven by gut feelings, but what about the claim that characteristically "utilitarian" moral judgments are driven by calculative reasoning? Researchers have found, as the model predicts, more "utilitarian" responses to dilemmas among people with emotional deficits, such as psychopaths, people suffering from frontotemporal dementia, and patients with damage to the vmPFC (Koenigs et al. 2012; Mendez et al. 2005; Koenigs et al. 2007). Activity in the amygdala correlates negatively with "utilitarian" judgments but positively with

adverse feelings in personal moral dilemmas (Shenhav & Greene 2014). Related to this finding, people with psychopathic traits appear to exhibit lower amygdala activity when responding to personal dilemmas (Glenn et al. 2009). As a final example, "utilitarian" responses are lower among participants whose brains had been flooded with serotonin, which especially influences the amygdala and vmPFC, among some other regions (Crockett et al. 2010).

Despite the array of corroborating evidence, there are many criticisms of the very dual-process elements of the theory. For example, the personal/impersonal distinction, based largely on reaction time data, was driven by a handful of stimuli from the complete set of about 60 dilemmas judged by participants (McGuire et al. 2009). Furthermore, while different reaction speeds—fast or slow—can reflect intuitive or deliberative processes, this may be due to the particular examples of those types of dilemmas the researchers happened to choose (Krajbich et al. 2015). Indeed, moral dilemmas can be constructed that yield "utilitarian" responses that are intuitive and "deontological" ones that are counter-intuitive (Kahane et al. 2012). Another concern is that the automatic versus controlled moral judgments measured with sacrificial dilemmas don't clearly track the relevant moral values or types of moral reasoning. Some apparently "utilitarian" resolutions to personal dilemmas appear driven by callousness (e.g. indifference to pushing), not a utilitarian concern for the greater good (Kahane et al. 2015; but see Conway et al. 2018).

Some versions of dual process theory also treat automatic moral intuitions as relatively simple and inflexible, which understates how they can be shaped by unconscious learning (Railton 2017). Indeed, some neuroimaging evidence suggests distinct brain areas underlie the resolution of moral dilemmas in terms of factors familiar from moral theory, such as the act/omission distinction and harming as a means versus a side effect (Schaich Borg et al. 2006). What seemed like a simplistic emotional aversion to pushing or prototypically violent acts in personal dilemmas turns out to be driven by complex concerns about how involved an agent is in bringing about harmful outcomes (Mikhail 2011; Feltz & May 2017; May 2018).

3.2 Beyond Dilemmas

Some neuroscientists have gone beyond sacrificial dilemmas and dual-process theory when investigating moral cognition (although they do remain fixated largely on harming/helping others). Most of the extant research has extensively studied moral judgments about hypothetical scenarios involving attempted versus accidental harms. Now, intentionality may not be crucial for all moral situations—sleeping with your cousin is deemed impure and morally problematic by many people, even if you're completely oblivious to the family connection (Young & Saxe 2011). Nevertheless, across cultures, an actor's mental states (intent, belief, knowledge, or lack thereof) influence moral evaluations of harmful acts (Barrett et al. 2016; McNamara et al. 2019).

Liane Young and her collaborators have found that increased activation in the TPJ/pSTS is associated with attribution of intent during the evaluation of attempted harms (Young et al. 2007; Young & Saxe 2008). One study even decoded

activity in this region—using multi-voxel pattern analysis—to predict individual differences in moral evaluations of accidental versus intentional harms (Koster-Hale et al. 2013). Moreover, while "No harm, no foul" usually doesn't apply to attempted murder, disrupting the TPJ with transcranial magnetic stimulation made participants judge less harshly an agent's failed attempt to harm someone, by downplaying the agent's malicious intent and focusing instead on the lack of harmful outcomes (Young et al. 2010).

These findings are consistent with theories suggesting that the TPJ—which overlaps with the pSTS—is critical for the domain-general capacity of mental state understanding and empathy (Decety & Lamm, 2007; Young & Dungan, 2012). Unsurprisingly, some evidence even suggests that patients with high-functioning autism have difficulty integrating mental state information into their moral judgments of harm, causing them to judge accidental harms more harshly than neurotypical individuals do (Moran et al. 2011). Similar results have been found with split-brain patients, which coheres with evidence that belief-attribution in the TPJ is partly lateralized to the right hemisphere (Miller et al. 2010).

A different neuroimaging paradigm asks participants to judge statements, instead of scenarios, as right or wrong, some of which go beyond harm and even fairness (Moll et al. 2001; Moll et al. 2002). While in the scanner, participants judged as "right" or "wrong" moral statements (e.g. "The elderly are useless"), non-moral statements (e.g. "The elderly sleep more at night"), and scrambled statements (e.g. "Sons push use eat work."). The researchers found that moral judgments, relative to judgments about non-moral statements, elicited greater activity in brain areas now familiar in moral neuroscience: vmPFC, left temporal pole (highly connected to the vmPFC and amygdala), and TPJ/pSTS.

Studies using electroencephalography (EEG) further suggest a temporal order over which various neural circuits contribute to the moral evaluation of harmful acts. When forming such moral judgments, participants rapidly computed information about mental states in TPJ, around 60 milliseconds after viewing a short video depicting accidental or intentional harm (Decety & Cacioppo 2012). Just a few hundred milliseconds later, the amygdala provided evaluative input to areas of the prefrontal cortex before a moral judgment emerged, whether concerning harmful or helpful acts (Yoder & Decety 2014). At least when it comes to the domain of harm, the brain appears to compute the positive and negative consequences of the act while weighing up how involved an agent was in bringing them about.

3.3 Limitations

One criticism facing much work in the second thread is the overreliance on "reverse inference" to infer the existence of certain mental states from activations in brain areas, when such areas perform multiple psychological functions (see e.g. Klein 2010). The amygdala, for example, is associated with among other things motivation, fear, and reward; greater activity in the amygdala when participants give "deontological" responses to moral dilemmas doesn't necessarily mean participants felt increased motivation as opposed to fear, reward, or perhaps

something else entirely by way of functional interactions with other regions at the network-level.

Reverse inference can be warranted, however, if the observation of brain activity in some region provides better support for one theory over another (Machery 2014). Moreover, as we've seen, moral neuroscience does not rely solely on neuroimaging but also on other methods that are less susceptible to concerns relating to reverse inference. We have seen some appeal to lesion studies, which go beyond merely correlating brain areas with moral cognition to provide evidence that a region is *necessary* for moral cognition. Some research is even able to discern which areas are necessary and *sufficient* for moral cognition by employing non-invasive brain stimulation techniques or psychotropic drugs (or both). Some of the studies cited above used transcranial magnetic stimulation, for instance, to increase (or decrease) neuronal excitation or medications to enhance (or impair) the functioning of neurochemicals such as serotonin. Even the dual-process model, despite being born of fMRI, has been tested against brain lesion data and the manipulation of neurotransmitters, not to mention various psychological experiments.

Another important limitation is that extant theories are woefully *incomplete*. Much of moral judgment doesn't involve death or bodily harm, let alone dilemmas featuring these moral considerations. Reams of evidence now suggest that across cultures fundamental moral values include not just harm or care but also fairness, group loyalty, sanctity, and respect for authority (Gilligan 1982; Haidt et al. 1993; Doğruyol et al. 2019). Even if these other moral values are ultimately reducible to harm/care, moral neuroscientists have largely ignored them (a few exceptions: Parkinson et al. 2011; Decety, Pape, & Workman, 2018; Workman, Yoder, & Decety 2019).

3.4 Lessons

The second thread in moral neuroscience primarily examined differences in brain activity elicited by moral compared to non-moral stimuli, or to moral stimuli of one kind compared to another, which were then localized to specific parts of the brain. Combined with lessons from the first thread, a general picture emerges in which at least some core aspects of moral cognition are underpinned by a network of predominantly frontal and temporal regions dedicated to various morally-relevant factors recognizable to both commonsense and moral theorizing (for further reviews, see Moll et al. 2005; Greene 2009; Eres et al. 2017; Han 2017; Demaree-Cotton & Kahane, 2018).

A central lesson is that moral cognition is not localized to one brain area or otherwise particularly unified in the brain (Greene 2009; Young & Dungan 2012; Parkinson et al. 2011). Instead, our capacity for moral judgment involves a spatially distributed network of areas with various domain-general psychological functions that are also relevant to moral evaluation, such as understanding the consequences of an agent's action, the agent's mental states, how the action was causally brought about, and the social norms it violates.

Another key lesson is that moral judgment is not always driven only by gut feelings or localized just to vmPFC and amygdala. In contrast with the first thread,

we see that some moral judgments involve rapid *reasoning*, served by areas such as dlPFC and TPJ. Moreover, even automatic moral intuitions can involve complex computation (Mikhail 2011). While some moral intuitions are heavily dependent on amygdala and vmPFC, these are part of a network of brain areas that engage in sophisticated, even if unconscious, learning and inference (Huebner 2015; Woodward 2016).

4. Third Thread: Learning

The first and second threads in moral neuroscience focused on brain areas and their functions. A third thread focuses on the level of computational analysis in neuroscience. In particular, many proponents of this type of approach draw on reinforcement learning to illuminate the nature and workings of moral judgment.

4.1 Value, Reward, and Learning

The field of reinforcement learning asks how an agent can learn to optimize its behavior strictly from interactions with its environment. For example, a baby plover is able to leave the nest and feed itself within a few hours of hatching, receiving relatively little assistance from its parents. How is it able to learn and perform these behaviors?

Research in reinforcement learning analyzes idealized versions of this question (Sutton & Barto, 2018). Specifically, it recasts the question in terms of how an agent can learn to maximize its reward and value over time. *Reward* refers to the intrinsic desirability of a given stimulus, whereas *value* refers to the total, expected, future reward associated with a given state. For example, drinking a cup of coffee is intrinsically desirable for many people because it is flavorful and provides caffeine, and so is rewarding. By contrast, grinding some coffee by hand is not rewarding, and may even be annoying, but it is valuable because it subsequently leads to the rewarding state. Many reinforcement learning methods use the notions of reward and value to estimate what it would be good for an agent to do in the long run.

The computational study of reward and value extends into the study of the neurobiological bases of human decision-making. An early discovery in computational neuroscience revealed an important correspondence between one of these reinforcement learning methods, known as the Temporal-Difference learning algorithms, and the firing of dopamine neurons in the mammalian brain (Schultz et al. 1997). A substantial body of animal and human behavioral evidence has since suggested that there are at least three different decision systems in the brain: the Pavlovian system, the model-free (or habitual) system, and the model-based (or goal-directed) system (e.g., see Glimcher and Fehr 2014).

The *Pavlovian* system produces basic, stimulus-driven behavioral responses. The term “Pavlovian” frequently leads to confusion (see Rescorla 1988). In most fields, as well as in everyday usage, the term is usually associated with Pavlov’s original experiments with dogs, where he conditioned dogs to salivate at the sound of a bell by repeatedly ringing a bell and then consistently feeding them

afterwards. By contrast, in reinforcement learning, "Pavlovian" refers to the relationship between the *unconditioned* stimulus (the food) and the relevant unconditioned response (the salivating). Thought to be evolved and roughly "hardwired," these unconditioned responses include both outcome-specific responses, such as inflexibly licking water, and more open-ended, valence-dependent responses, such as generally approaching something rewarding. Both classes of response are characteristically recalcitrant to changes in outcome, as when chickens will continue to peck at a feeder that will not dispense any seeds over hundreds of trials (Macintosh, 1983; Huys et al. 2011, 2012). The Pavlovian system is supported primarily by the brain stem and subcortical areas of the limbic system—e.g. amygdala, nucleus accumbens, and hypothalamus (Rangel et al. 2008).

The *model-free* system produces instrumental responses by evaluating actions based on their previously learned values in different contexts. For example, a button-press that has previously resulted in a reward is a good state-action pair, while a button-press that has previously resulted in a punishment is a bad state-action pair. Because the model-free system does not explicitly represent future values, it can be slow to update in the face of changing circumstances. However, unlike the Pavlovian system, the model free system is not "hardwired," and does gradually update. The model-free system is associated with activity in the basal ganglia and the orbital and medial portions of the prefrontal cortex (Yin and Knowlton, 2006).

Finally, the *model-based* learning system uses a forward-looking model to represent possible actions, outcomes, and associated values. This model is typically represented by a *decision tree*. Each node in the tree represents a possible choice, where the model-based system "searches" through the decision tree to find the branch with the highest total value. For example, a chess player may represent three upcoming moves in a game of chess, with each possible move further branching into a wide range of subsequent moves. To win, the player tries to represent and choose the best possible sequence of moves overall. The model-based system is primarily associated with activation in the vmPFC (Hare et al. 2008).

4.2 Moral Learning

Learning-based approaches to moral judgment are developed using the three decision-systems. Which system plays a defining role in moral judgment? Echoing the dual-process theories discussed in the second thread, Cushman (2013, 2015) argues that much of moral cognition depends on a body of objective rules together with the model-free decision system (see also Greene 2017). When exhibiting the latter process, people often continue to adhere to norms outside of the context in which those norms are in play. For instance, American tourists frequently continue to tip in restaurants abroad, even when there is no relevant norm dictating that they should (2015, 59).

Cushman argues that the role of the model-free decision system helps explain participants' diverging responses to the Switch and Footbridge scenarios. Cushman reasons that people's tendency to resist harming the single individual in

Footbridge is "the consequence of negative value assigned intrinsically to an action: direct, physical harm" (2015, 59). That is, participants' responses may be underwritten by the model-free decision-system: since directly harming others has reliably elicited punishments in the past, this option represents a bad state-action pair, and leads people to reject it as an appropriate course of action.

One difficulty with Cushman's general view is that it is in tension with the aforementioned evidence suggesting that all three decision systems trade off and interact to produce our everyday behaviors. Another, more specific difficulty comes from the fact that participants' avoidance of harm can just as plausibly be explained by the role of the evolved, Pavlovian system as can by the role of its model-free counterpart. One way to disentangle which system is in effect could be to devise an iterative version of the trolley problem. If participants gradually shifted their views on the matter, we could say that it was the model-free system; if they did not, we could say that it was recalcitrant Pavlovian responding.

In contrast to Cushman's highlighting of the role of only one of the three decision systems in producing moral judgments, Crockett (2013, 2016) argues that all three systems play a role, and even interact in the process of producing a single judgment. On Crockett's view, *both* the model-free and Pavlovian systems assign negative values to actions that cause others harm through physical contact. Consequently, when the "votes" of all three systems are tallied up, participants will find it morally acceptable to sacrifice one to save five in impersonal dilemmas, but not in personal dilemmas. Hence, even the iterative version of Cushman's view and Crockett's voting explanation provide competing explanations of responses to the trolley problem, and so leave open questions for further investigation.

4.3 Lessons

Computational approaches to understanding moral judgment complement rather than compete with the first and second threads discussed above. In particular, computational approaches complement strictly behavioral and neuroscientific accounts by illuminating the relationships between the components of moral cognition, using formal mathematical models. Adopting such strategies has the further advantage of enabling researchers to leverage additional bodies of research from computer science and economics.

Notably, this third thread in the neuroscience of morality coheres with the suggestion commonly found in other threads that we use domain-general decision-making systems to make specifically moral judgments. It seems we use the same algorithms and neural mechanisms to make, for example, choices about which car to buy and decisions about which people to blame for moral wrongdoing (see Berns et al. 2012; Shenhav and Greene 2010; Crockett et al. 2017). This emerging picture is also consistent with findings in the first thread which suggest that breakdowns in general decision-making are associated with related breakdowns in moral decision-making (Blair et al. 2001; Mahmut et al. 2008; Aharoni et al. 2012).

Limitations found in previous threads, however, remain, including the ongoing reliance on sacrificial dilemmas. Going forward, computational approaches will need to model more than judgments about dilemmas and go beyond

the domain of harm by studying loyalty, care, and other values that arise in everyday circumstances.

5. Future Directions: Moral Development

Moral learning is a process that occurs over time and stretches back to the critical period of childhood. We saw in Section 2 that psychopathy involves dysfunction of at least amygdala and vmPFC, presumably early in development (Glenn & Raine 2014; Taber-Thomas et al. 2014), which seems to affect one's moral capacities. In contrast, moral deficits are much less profound in patients who acquire damage to these regions in adulthood, because normal brain development affords the acquisition of moral competence. Psychopathy is only one form of moral dysfunction, however, and we should seek a complete understanding of normal moral development that incorporates not only empathy and compassion but also harm (including the weighing of outcomes and the actor's intent) and other moral values (e.g. loyalty). Neuroscientists are increasingly interested in understanding how such elements of moral cognition work and develop in the brains of children and adolescents.

5.1 Moral Judgment and Brain Development

As in other areas of neuroscience, we do well to consider brain development in conjunction with relevant psychological theories and evidence. Building on work by Piaget (1932), Kohlberg and his colleagues theorized the development of moral cognition in terms of reasoning and reflection that, once fully developed, employed universal principles that could even be used to question existing conventions in society (Kohlberg 1984).

An important concern, however, is that this approach only tracks the development of conscious moral reasoning, which could be merely rationalizing moral judgments one has already made intuitively on different grounds (Haidt 2001). If we seek a theory of moral judgment, not of our (often poor) attempts to justify them verbally, then we need to explain the development of unconscious processes that generate automatic moral intuitions (Cushman et al. 2006).

Taking this approach, researchers have investigated moral development with age-appropriate moral scenarios. Using morality plays with puppets, for example, researchers have found that even infants discriminate and prefer a puppet that help other characters achieve, as opposed to hinder, their goals (Hamlin 2015; Cowell & Decety 2015). Children as young as four begin making moral judgments focused on outcomes, such as whether an action harmed or saved more people (Pellizzoni et al. 2010), regardless of whether it was accidental or intentional. The intent of the actor appears to grow increasingly relevant in the next few years of development (Cushman et al. 2013).

Corresponding to the psychological research, studies in developmental neuroscience have found relevant differences in brain structure and function across age groups during moral cognition. One neuroimaging study, for example, found greater activity in vmPFC in adults, compared to younger participants, when they viewed moral relative to non-moral transgressions (Decety, Michalska, & Kinzler

2012). Among older participants, the researchers also observed greater task-based functional connectivity between the vmPFC and amygdala, and between the vmPFC and TPJ/pSTS. Consistent with other studies, younger participants' evaluations of a person who caused harm were less sensitive to whether the harm was intentional or accidental. In another study, when both adolescent and adult males evaluated images of moral violations, researchers found greater activity in the TPJ/pSTS and PCC among older participants (Harenski et al. 2012). In their review of these and other studies, Decety and Cowell conclude that "mature moral cognition" at least requires continued development in brain areas that underlie "aversion to interpersonal harm, empathic concern, and mental state understanding" (2018, 160).

Further research is needed, but studies in developmental neuroscience thus far fit well with the moral circuits identified in adulthood. Central players include limbic regions (particularly the amygdala), portions of the prefrontal cortex (especially vmPFC), and relevant areas of the temporal lobe (namely, the STS, including its posterior/TPJ). Brain activity in these moral circuits changes over the course of development, and such changes are associated with key elements of moral cognition, particularly: assigning value to outcomes such as harm, representation of the actor's knowledge or intentions, and retrieval of relevant social information. However, again, morality involves more than harm or even fairness (Gilligan 1982; Haidt et al. 1993; Doğruyol et al. 2019). Further developmental neuroscience research should study more than simplistic depictions of harm, altruism, or compassion and make sure their findings generalize to, say, fraud, torture, betrayal, and filial duties.

5.2 Integrative Approaches

The neuroscience of moral development also suggests an interesting overlap between regions that support moral cognition and regions that support thinking about the self. Meta-analyses of neuroimaging studies reveal that several moral circuits—e.g. the vmPFC and PCC—overlap with the default mode network (Bzdok et al. 2012; Eres et al. 2017; Sevinc & Spreng 2014; Han 2017). In many studies, participants are asked to evaluate *other* people and their actions, so it's striking to find such extensive overlap with self-related regions.

One explanation for this is that participants often make moral judgments in response to emotionally-charged stories with actors who intend to cause harm, which naturally recruits brain areas that contribute to understanding narratives, theory of mind, and distinguishing self from other. However, another explanation is that, while moral judgment and motivation are distinguishable, they are intimately connected, especially in normal development. Extensive interview studies do suggest the integration of moral values and one's self-concept occurs throughout adolescence and into adulthood (Damon 1984). Moreover, we've seen that psychopathy affects moral cognition not only by causing dysfunction in areas associated with conscious reasoning or social knowledge, but emotion and motivation.

Thus, it may be that a brain develops normal moral judgment only through proper development of a suite of connected moral capacities, including emotions,

motivation, and identity. An analogy may help. Suppose that in educated adults the ability to solve algebraic equations is localized, more or less, to the prefrontal cortex. It doesn't follow that a child can *learn* algebra so long as the prefrontal cortex is functioning properly. If other areas are dysfunctional during development—even including swaths of motor or primary visual cortex—one may be unable to properly develop their mathematical capacity, even if this is later grounded in only one of the many brain areas necessary for initial development.

One approach would be to integrate moral judgment, motivation, and behavior via *moral identity*, or the degree to which moral values are central to one's self-concept (Aquino & Reed II 2002). Experimental evidence suggests that people are more likely to behave according to their moral judgments if they regard moral values as both central to themselves and more important than non-moral values (Reed et al. 2007; Winterich et al. 2013). Qualitative studies corroborate the idea that strong moral identity is required for sustained commitment to moral behavior (Colby & Damon 1992). Psychopathy may even involve weak moral identity, since people with psychopathic tendencies have been shown to report weaker moral identities (Glenn et al. 2010).

Another integrative approach, which is rather mainstream in moral education (Han 2014), is neo-Kohlbergian. Unlike classical Kohlbergian theory, which focused only on moral judgment and reasoning, the Four Components Model incorporates additional aspects of moral functioning (Bebeau 2002)—namely, moral motivation, character, and sensitivity. On this model, moral development and maintenance involves orchestrating these four components to cooperate with each other (Rest & Narvaez 1994), which are associated with interactions among various limbic and frontal regions (Narvaez & Vaydich 2008).

We can perhaps situate integrative developmental theories within integrative models of the neurobiology of *mature* moral judgment. The Event-Feature-Emotion framework (Moll et al. 2005), for example, identifies a spatially distributed network of frontal, temporal, and subcortical brain areas involved not just in the moral evaluation of others but also in moral emotion and motivation. Such frameworks cohere with the meta-analysis of neuroimaging studies suggesting that moral circuits significantly overlap with self-related psychological processing (Han 2017). Thus, when it comes to the development of moral cognition and its improvement in adulthood, it is wise to consider the integration of otherwise dissociable moral capacities, including both moral judgment and motivation (May 2018).

6. Conclusion: Philosophical Implications

The neuroscience of moral judgment is still fairly young. There is no doubt that conclusions about moral judgment on the basis of neurobiology should be drawn with caution. Nevertheless, in this final section, we aim to show how combining brain science with philosophical analysis can aid our understanding of moral judgment, particularly by elucidating concrete mechanisms and corroborating or disconfirming theories developed in other areas of cognitive science (Prinz 2016; Demaree-Cotton & Kahane 2018).

Indeed, the neuroscience of ethics is already greatly improving, due to philosophy and science continuously informing one another. We've already seen how decades of ethical theorizing about the trolley problem, for instance, has shaped experimental paradigms. In this section, however, let's conclude by briefly drawing out how the advances in neuroscience discussed above can contribute to debates in moral philosophy.

6.1 Reason vs. Emotion in Ethics

The dichotomy between reason and emotion stretches back to antiquity. But an improved understanding of the brain has, arguably more than psychological science, questioned the dichotomy (Huebner 2015; Woodward 2016). Brain areas associated with prototypical emotions, such as vmPFC and amygdala, are also necessary for complex learning and inference, even if largely automatic and unconscious. Even psychopaths, often painted as the archetype of emotionless moral monsters, have serious deficits in learning and inference. Moreover, even if our various moral judgments about trolley problems, harmless taboo violations, and the like are often automatic, they are nonetheless acquired through sophisticated learning mechanisms that are responsive to morally-relevant reasons (Railton 2017; Stanley et al. 2019). Indeed, normal moral judgment often involves gut feelings being attuned to relevant experience and made consistent with our web of moral beliefs (May & Kumar 2018).

Blurring the line between reason and emotion may seem to render the corresponding philosophical disputes meaningless, but that's too fast. If emotions are required for moral judgment only because affect is integral to reasoning generally, then moral judgment isn't special in requiring emotional processes, which is a core tenant of sentimentalism. Instead, what seems vindicated is a core thesis of rationalism: that moral cognition involves domain-general capacities for learning and reasoning, just like non-moral cognition. Rather than obliterate the philosophical dispute, the evidence may support sophisticated form of rationalism (May 2018), despite early proclamations that the science preferentially supports sentimentalism.

6.2 Moral Knowledge (or Lack Thereof)

A consensus already seems to be emerging that, under the skull's hood, moral cognition is a complex affair, not just among ethicists and the intelligentsia but also ordinary people. Moral judgment is not merely a matter of expressing one's emotions, divorced from reasoning. Moralizing is informed at least by one's own experiences as well as knowledge from one's society and ancestors, in the form of cultural norms and evolved predispositions. Yet moral beliefs are not fixed after maturation. Even staunchly held moral attitudes, such as opposition to same-sex marriage, can rapidly change in response to greater understanding of others and consistency reasoning (Campbell & Kumar 2012).

However, even if most moral cognition involves learning and inference, these may be too biased and unreliable to yield moral knowledge or justified belief. By helping to uncover the causal sources of moral cognition, neuroscience can aid in either debunking or vindicating certain kinds of moral beliefs (Greene 2017;

Kumar 2017), although sweeping conclusions about all moral cognition are likely to falter (May 2018; Kumar & May 2019). Of course, neuroscience alone can't settle philosophical issues without making normative assumptions (Berker 2009), but together they can advance debates in ethics.

6.3 Moral Education and Progress

Understanding the workings of mature moral judgment, as well as its development, also promises to illuminate how we can improve the acquisition of moral knowledge and perhaps even speed up moral progress. Extant evidence already suggests that mature moral judgment requires the proper development of an interwoven tapestry of moral capacities, including appropriate reasoning, sentiments, motivations, learning mechanisms, and self-conception.

Of course, neuroscience alone is unlikely to demonstrate how to improve our moral selves. But such findings *can* suggest useful directions for moral psychology and moral education, especially when applied to a particular social context. For example, given the association between morality and identity at the neural level, Han and his colleagues (2017) predicted and found that stories of closely related moral exemplars, such as peers, more effectively promoted moral elevation and emulation than stories of distant exemplars, such as historical figures (see also Han, Workman, Dawson, & May 2018). Or consider more newfangled proposals for moral improvement, such as indiscriminately amplifying moral emotions—whether through pills, brain stimulation, or lacing the water with oxytocin (Earp, Douglas, & Savulescu 2017). The neuroscience of moral judgment already speaks against such tactless tactics. Of course, devastating a person's moral capacities may be as simple as disrupting moral circuits in childhood—unfortunately, it's generally easier to harm than to benefit (Persson & Savulescu 2012). But distinguishing right from wrong is an incredibly complex process that requires the coordinated orchestration of a diverse range of brain areas and neurotransmitters. Novel neurobiological methods for moral improvement will certainly require finesse.

Acknowledgements

For helpful feedback on this chapter, we thank the editors and other contributors to this volume, as well as Vanessa Bentley, Jean Decety, Andrea Glenn, and Andrew Morgan. Special thanks to Felipe De Brigard and Walter Sinnott-Armstrong for directing the Summer Seminars in Neuroscience and Philosophy and to the John Templeton Foundation for funding them. Authorship: JH wrote Section 4; HH wrote section 5; JM and CW co-authored the remaining sections and edited the entire manuscript.

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.

Allen, C., & Wallach, W. (2009). *Moral Machines: Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

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.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of personality and social psychology*, 83(6), 1423–1440.

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.

Bebeau, M. J. (2002). The defining issues test and the four component model: contributions to professional education. *Journal of moral education*, 31, 271–295.

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

Berns, G. S., Bell, E., Capra, C. M., Prietula, M. J., Moore, S., Anderson, B., ... & Atran, S. (2012). The price of your soul: neural evidence for the non-utilitarian representation of sacred values. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 754-762.

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

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, 217, 783–796.

Campbell, R., & Kumar, V. (2012). Moral Reasoning on the Ground. *Ethics*, 122(2), 273–312.

Colby, A., & Damon, W. (1992). *Some do care : contemporary lives of moral commitment*. New York, NY: Free Press.

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

Cowell, J. M., & Decety, J. (2015). Precursors to Morality in Development as a Complex Interplay between Neural, Socioenvironmental, and Behavioral Facets. *Proceedings of the National Academy of Sciences* 112(41): 12657–12662.

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

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

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107(40), 17433–17438.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature neuroscience*, *20*(6), 879.

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

Cushman, F. (2015). From moral concern to moral constraint. *Current opinion in behavioral sciences*, *3*, 58-62.

Cushman, F. A., Young, L. L., & Hauser, M. D. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm, *Psychological Science*, 17(12), 1082–1089.

Cushman, F. A., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, 127(1), 6–21.

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

Damon, W. (1984). Self-understanding and moral development from childhood to adolescence. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Morality, moral behavior and moral development* (pp. 109–127). New York, NY: John Wiley & Sons.

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

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology*, 30(1), 153–164.

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.

Decety, J., Pape, R., & Workman, C. I. (2018). A multilevel social neuroscience perspective on radicalization and terrorism. *Social Neuroscience*, 13(5), 511–529.

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. *Cerebral cortex* 22(1), 209–20.

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.

Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, 151, 109547.

Earp, B. D., Douglas, T., & Savulescu, J. (2017). Moral neuroenhancement. In *The Routledge Handbook of Neuroethics*, L. S. Johnson, K. S. Rommelfanger (eds.). Routledge, pp. 166-184.

Eres, R., Louis, W. R., & Molenberghs, P. (2017). Common and distinct neural networks involved in fMRI studies investigating morality: an ALE meta-analysis. *Social Neuroscience*, 1–15.

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

Gazzaniga, M. S. (1983). Right Hemisphere Language Following Brain Bisection. *American Psychologist*, 38(5), 525–537.

Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Harvard University Press.

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, 5(7), 497–505.

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

Glenn, A. L., Raine, A., & Schug, R. A. (2009). The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry*, 14(1), 5–6.

Glimcher, P. W., & Fehr, E. (Eds.). (2014). *Neuroeconomics: Decision making and the brain*. New York: Academic Press.

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

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.

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

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog. *Journal of Personality and Social Psychology*, 65(4), 613–628.

Hamlin, J. K. (2015). The Infantile Origins of Our Moral Brains. *The Moral Brain*, J. Decety and T. Wheatley (eds.). MIT Press, pp. 105-122.

Han, H. (2014). Analysing Theoretical Frameworks of Moral Education through Lakatos's Philosophy of Science. *Journal of Moral Education*, 43(1), 32–53.

Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: A meta-analysis. *Journal of Moral Education*, 46(2), 97–113.

Han, H., Kim, J., Jeong, C., & Cohen, G. L. (2017). Attainable and Relevant Moral Exemplars Are More Effective than Extraordinary Exemplars in Promoting Voluntary Service Engagement. *Frontiers in Psychology*, 8, 283.

Han, H., Workman, C., Dawson, K. J., & May, J. (2018) "Which Stories of Moral Exemplars Best Provoke Moral Behavior?" Presentation at the 44th Annual Conference of Association for Moral Education, Barcelona, Spain.

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.

Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2012). Neural development of mentalizing in moral judgment from adolescence to adulthood. *Developmental Cognitive Neuroscience* 2, 162–173.

Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In *Philosophy and Computing* (pp. 121-159). Springer, Cham.

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

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.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS computational biology*, *7*(4), e1002028.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

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

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.

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

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.

Klein, C. (2010). Philosophical Issues in Neuroimaging. *Philosophy Compass*, 5(2), 186–198.

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.

Kohlberg, L. (1984). *The psychology of moral development: the nature and validity of moral stages*. San Francisco: Harper & Row.

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.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, 6(1), 7455.

Kumar, V. (2017). Moral Vindications. *Cognition*, 167(Oct), 124–134.

Kumar, V. & May, J. (2019). How to Debunk Moral Beliefs. J. Suikkanen & A. Kauppinen (eds.), *The New Methods of Ethics*.

Landy, J. F., & Goodwin, G. P. (2015). Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence, *Perspectives on Psychological Science*, 10(4), 518–536.

Macintosh, N.J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

Macmillan, M. (2000). *An Odd Kind of Fame: Stories of Phineas Gage*. MIT Press.

Machery, E. (2014). In Defense of Reverse Inference. *The British Journal for the Philosophy of Science*, 65(2), 251–267.

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

Mahmut, M. K., Homewood, J., & Stevenson, R. J. (2008). The characteristics of non-criminals with high psychopathy traits: Are they similar to criminal psychopaths?. *Journal of Research in Personality*, *42*(3), 679-692.

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

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

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, 45(3), 577–580.

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, M. B., Sinnott-Armstrong, W., Young, L. L., King, D., Paggi, A., Fabri, M., et al. (2010). Abnormal moral reasoning in complete and partial callosotomy patients. *Neuropsychologia*, 48(7), 2215–2220.

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.

Narvaez, D., & Vaydich, J. L. (2008). Moral Development and Behaviour Under the Spotlight of the Neurobiological Sciences. *Journal of Moral Education*, 37(3), 289–312.

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

Padoa-Schioppa, C., & Schoenbaum, G. (2015). Dialogue on economic choice, learning theory, and neuronal representations. *Current opinion in behavioral sciences*, *5*, 16-23.

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.

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science*, 13(2), 265–270.

Persson, I., & Savulescu, J. 2012. *Unfit for the future: The need for moral enhancement*. Oxford University Press.

Piaget, J. (1932). *The Moral Judgment of the Child*. Free Press.

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.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7), 545-556.

Reed, A., Aquino, K., & Levy, E. (2007). Moral Identity and Judgments of Charitable Behaviors. *Journal of Marketing* 71(1), 178-193.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist* 43(3): 151-160.

Rest, J. R., & Narvaez, D. (1994). *Moral development in the professions: Psychology and applied ethics*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Robertson, D., Snarey, J., Ousley, O., Harenski, K., Bowman, F. D., Gilkey, R., & Kilts, C. (2007). The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia*, 45(4), 755-766.

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.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of cognitive neuroscience*, 18(5), 803-817.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Sevinc, G., & Spreng, R. N. (2014). Contextual and perceptual brain processes underlying moral cognition: A quantitative meta-analysis of moral reasoning and moral emotions. *PLoS ONE*, 9, e87427.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

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.

Stanley, M. L., Yin, S., & Sinnott-Armstrong, W. (2019). A Reason-Based Explanation for Moral Dumbfounding. *Judgment and Decision Making*, 14(2), 120–129.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

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.

Winterich, K. P., Aquino, K., Mittal, V., & Swartz, R. (2013). When moral identity symbolization motivates prosocial behavior: the role of recognition and moral identity internalization. *Journal of Applied Psychology*, 98(5), 759-770

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.

Workman, C. I., Yoder, K. J., & Decety, J. (2019). The Dark Side of Morality: How Moral Convictions Facilitate Support for Political Violence. Manuscript under review.

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

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464.

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.