Synthese (2018) 195:1873–1889
https://doi.org/10.1007/s11229-017-1344-9

S.I.: NEUROSCIENCE AND ITS PHILOSOPHY

A neuropsychological challenge to the sentimentalism/rationalism distinction

Geoffrey S. Holtzman[1]

Received: 31 May 2016 / Accepted: 18 February 2017 / Published online: 17 March 2017

Abstract Critical reflection on the available neuropsychological evidence suggests that the roles of emotion and reason in moral judgment may not be distinct. This casts significant doubt on our current understanding of moral judgment, and therefore also on all philosophical theories based on that understanding. Most notably, it raises doubts about both sentimentalism and rationalism, which historically have often been treated as exclusive and exhaustive theories regarding the nature of moral concepts. As an alternative, I endorse pluralism with regard to the emotional and rational nature of moral concepts.

Keywords Sentimentalism · Rationalism · Pluralism · Emotion · Reason · Cognition · Neuroethics · Moral psychology · Dual-process model · Joshua Greene · Moral judgment task · Dilemmas · Trolley problems

The present essay was born of a desire to understand how we might best characterize the neural substrates involved in the processing of moral vignettes.[1] Does the currently dominant emotion/cognition dual-process model (Sect. 1)—which corresponds nicely to the sentimentalist/rationalist debate in moral philosophy—fit our most recent findings in the neuroscience of moral processing (Sect. 2)? I argue that it does not, and that we are therefore faced with two questions. First, why have we been so quick to accept that this fit exists if in fact it does not (Sect. 3)? And second, how might we

[1] Although Greene and others typically use the term 'dilemmas,' I will use the terms 'vignettes' or 'problems' except when writing quotationally. This is because, as Berker (2009) has pointed out, several of the vignettes used in Greene et al.'s (2001a) battery do not actually seem to involve genuine dilemmas.

✉ Geoffrey S. Holtzman
geoffreyholtzman@gmail.com

[1] Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL, USA

better describe the way the human brain identifies and processes conflicting ethical considerations, and (often) produces univocal responses in light of those considerations (Sect. 4)? After addressing these questions, I synthesize the previous sections to discuss future directions, and to draw two main conclusions (Sect. 5).

First, philosophers would do well to trade away the aim of a concrete but ultimately inaccurate psychological model of moral judgment in favor of a more philosophically rigorous, if perhaps less elegant, pluralistic account of moral processing. Although psychologists and neuroscientists may be best off embracing a clear, straightforward working model of the relationship between neural processing and moral judgment, philosophers are more likely to be led astray by such a model, given that the actual relationship between the two is neither clear nor straightforward. Thus, the kind of pluralism I am advocating is not necessarily concerned with pluralism in the normative criteria for what counts as (im)moral. Rather, my account is pluralistic specifically in the connections it draws between neuropsychological processes on the one hand, and moral judgments on the other.

Second, moral neuropsychology cannot be accurately characterized as a battle between emotion and controlled cognition, nor as a battle between sentiment and reason. Thus, researchers should reconsider the feasibility of characterizing the neural substrates of moral processing according to either sentimentalist or rationalist accounts.[2] For even if it turns out that sentimentalism or rationalism correctly describes how we *ought* to make moral judgments, it by no means follows that these accounts are helpful in understanding how we actually *do* make such judgments.

1 Background

Greene et al. (2001a; 2004) argue on the basis of fMRI data that moral vignettes[3] involving so-called 'personal' harms (ostensibly, those involving gratuitous physical harm to a specific person) tend to elicit automatic processes such as emotion, which in turn underwrite deontological moral judgments. In contrast, they argue, 'impersonal' moral vignettes (ostensibly, those that do not involve gratuitous physical harm or which are framed abstractly) tend to utilize working memory, and to activate more deliberative processes of reason and controlled cognition, leading to more utilitarian judgments.

The present essay will confine itself to discussing the roles of emotion and reason in moral judgment without respect to any particular normative ethical theory. This is because, as far as I can tell, Greene et al.'s (2001a; 2004) data have little to do with characteristically utilitarian or deontological judgments, even according to Greene et al.'s own definitions of such judgments. For instance, Greene et al. define utilitarian judgments as those that seek to "maximize aggregate welfare" and to "serve the greater

[2] Roughly, sentimentalism is the view that (normative) moral judgments are (and should be) driven primarily by emotion; rationalism is the view that (normative) moral judgments are (and should be) controlled primarily by reason.

[3] These 60 vignettes (Greene et al. 2001b) have received significant criticism in the literature. They are often described as "trolley problems" (Foot 1967) despite the fact that many of them are quite unlike trolley problems (Berker 2009); appear to be unrealistic or contrived (Gray and Schein 2012); and fail to control for a number of confounds (Kamm 2009).

good" (2004). But killing your grandmother by poisoning her tea just "to play a trick on her" (Greene et al. 2001b) does not "maximize aggregate welfare." Nor does it "serve the greater good" to murder your boss just "to get him out of your life," to hire a man to rape your wife just "so that she will appreciate you as you comfort her," or to throw your unwanted baby in a dumpster just "in order to move on with your life" (Greene et al. 2001b).

Greene et al.'s experimental protocol (2001b) was designed only to manipulate whether actions caused gratuitous physical harm to a specific person.[4] This may have been sufficient for assigning the casual labels of 'personal' and 'impersonal' to certain actions,[5] but it does not enable one to draw sound conclusions about whether approbation or disapprobation of those actions constitutes deontological or utilitarian judgment. For the purposes of their (2004) study using that same protocol (and indeed, for most analyses of data collected using their protocol), they (and dozens of other researchers)[6] simply labeled all approbations of personal moral violations as utilitarian judgments, failing to attend to the fact that many of these violations do not serve the greater good.

What has ensued, then, could be construed as two separate debates that have often intersected. The first of these, which has received the overwhelming share of attention in the philosophical community, concerns the (emotional vs. rational) origins of deontology and utilitarianism. I view this debate as fruitless, since it arises from a characterization of moral judgments that is superficial and inaccurate in the ways just described. I will therefore not concern myself with claims about deontological and utilitarian judgment in this essay.

The second debate, which has received far less attention (but which has been the focus of some of Greene's more recent work), regards the claim that 'personal' moral vignettes tend to elicit emotional responses, whereas 'impersonal' vignettes tend to elicit controlled cognitive responses. It is this second debate that this essay is meant to engage and, ultimately, deflate. Participants in this second debate, as well as those in the first debate, have made two assumptions. While each of these assumptions may *prima facie* seem obviously true, both are undermined by recent discoveries in neuroscience.

The first assumption is that emotion and controlled cognition are distinguishable processes with distinctly identifiable neural underpinnings. The second assumption is that in determining moral judgment, emotion and cognition typically function antagonistically, i.e., in competition with one another. My purpose in this essay is to show that neither of these assumptions is warranted, and that contemporary neuroscientific evidence—including some of the very evidence cited by Greene et al. (2001a,

[4] Vignettes were graded by two independent coders according to three criteria: "The moral dilemmas of which the coders said that the action in question (a) could reasonably be expected to lead to serious bodily harm (b) to a particular person or a member or members of a particular group of people (c) where this harm is not the result of deflecting an existing threat onto a different party were assigned to the 'moral-personal' condition; the others were assigned to the 'moral-impersonal' condition" (Greene et al. 2001a).

[5] Although even this much may be doubted. Michael Davis has pointed out to me that, e.g., hiring a black-market surgeon to "kidnap a randomly selected stranger, carve out one of his eyes, and transfer it to you" (Greene et al. 2001b, "Impersonal Dilemmas" #19) is a personal moral violation even according to Greene et al.'s coding conventions.

[6] For examples, see reviews by Christensen and Gomila (2012) and Garrigan et al. (2016).

2004)—undermines both assumptions. Perhaps philosophers have only accepted these assumptions because our long traditions of sentimentalism and rationalism have led us to find the emotion/cognition dichotomy intuitive. But intuitions, as Greene himself likes to point out, can be misleading.

1.1 Problems with emotion/cognition dual-process models

Describing neural structures as "emotional"—as some philosophers (e.g., Kamm 2009) recently have—seems at best like a category mistake, and at worst just plain wrong. Greene himself (2014b) now thinks that the regions he initially associated with 'personal' moral judgment can no longer be properly characterized as the neural substrates of emotion. Rather, they appear to be parts of the so-called "default mode network," which is involved in numerous mental functions in the absence of external stimuli. These functions include autobiographical memory, spontaneous cognition, projections of the future, and the representation of narrative episodes (Buckner et al. 2008).

In fact, even describing these structures as *associated with emotion*, while perfectly correct, can be problematic. Such descriptions are frequently misinterpreted by philosophers to imply that these structures have been *dissociated from controlled cognition* (Sect. 3.1). But as I will explain (Sect. 2.1), regions associated with emotion are often also associated with controlled cognition—and with functions that cannot be classified as either emotional or cognitive.

Furthermore, the fact that a given brain region receives increased blood flow when a person considers certain non-moral problems that require controlled cognition (Vincent et al. 2008) does not, as philosophers such as Kamm seem to believe, make that region a "rational center of the brain" (2009). Nor does its increased activity in the face of certain other, moral problems reveal those moral problems to be resolved "rationally," or primarily on the basis of controlled cognition (Sect. 3.1). And evidence of the region's involvement among subjects answering a particular moral question certainly does not prove, as Peter Singer claims it does, that "the answer these subjects g[i]ve is, surely, the rational answer" (2005).

Assertions like Singer's assume that cerebral blood flow to specific brain regions correlates predictably with rationality. However, nothing in contemporary neuroscience justifies such a claim. Furthermore, even if we could detect *rationality* from blood flow, inferences like Singer's would "surely" be deductively valid only if *irrationality* could be dissociated from a given brain area (Sect. 3.1; Van Orden et al. 2001). Yet with rare exceptions (e.g., Klein 2011; Helion and Pizarro 2015), moral psychologists and moral philosophers have failed to devote sufficient scrutiny to the descriptive accuracy of labeling brain areas as rational or emotional.

1.2 A word on Joshua Greene

This essay is not meant to be an indictment of Joshua Greene or his work. Rather, it is meant to cast doubt on a great deal of work that has built upon his projects and also, paradoxically, on a great deal of work attacking his projects.

Greene's decision to contrast moral neuropsychological processes along the lines of emotion and reason is deeply rooted in philosophical tradition. Contemporary Western

ethical thought is grounded in early debates about the primacy of emotion versus reason in moral judgment. Kant (1964/1785) rooted his moral philosophy firmly in rationality; Hume, in contrast, thought that "reason is and ought only to be the slave of the passions" (2013/1751). Thus, Greene cannot be blamed for working within an emotion-versus-reason psychological paradigm. Additionally, much of the neuroscientific research described in this essay was conducted after Greene et al. (2001a) defended their model; for not being aware of this later research, Greene of course also cannot be faulted.

1.3 Have we been building on a shaky foundation?

Philosophers who have objected to Greene's overall conclusions, but who have nonetheless tried to build on the foundations of his work, have often wasted energy by building on precisely those elements that are least instructive from a philosophical standpoint. These are elements that Greene and his colleagues themselves say have "proven useful" in generating neuropsychological results, but are "by no means definitive" (2001a).

In the next section, I discuss current neuroscientific understanding of the brain regions mentioned in Greene et al.'s (2001a) paper. There are many other brain regions I could discuss here, but I have two reasons for focusing only on these. First, while the neuroscience of moral judgment has evolved significantly over the years, commentary on this work continues to fixate on Greene's 2001 and, to a lesser extent, 2004 collaborative work.[7] Second, and perhaps more importantly, discussing any more regions than Greene et al. (2001a) mention would be superfluous.

In shadowing the philosophical literature, I focus entirely on neuroimaging data that have been brought to bear on the dual-process model of moral judgment and, in turn, on debates between sentimentalists and rationalists. By narrowing my focus in this way, I will be forced to overlook more recent behavioral studies of healthy, brain-damaged, and medicated persons. Those interested in such studies should consult Greene (2014a).

2 Re-examining brain areas associated with moral judgment

In this section, I demonstrate that it is inaccurate for philosophers to characterize the brain regions involved in different kinds of moral judgment as distinctly emotional or distinctly cognitive. Therefore, it is pointless to debate the proper way to interpret the normative or metaethical significance of the involvement of brain regions characterized as such.

2.1 Areas associated with emotion

Greene et al. (2001a) reported four brain regions to be activated more by 'personal' than by 'impersonal' moral vignettes. On the grounds that "recent functional imaging

[7] During 2015, all seven of Greene's solo and collaborative 2014 publications combined were cited less frequently (103 citations; M = 14.7) than either his 2001 (251 citations) or 2004 (172 citations) collaborations.

studies have associated each of these areas with emotion" (Greene et al. 2001a), they argue that "social-emotional responses […] undergird the absolute prohibitions that are central to" a distinct class of moral judgments[8] (Greene et al. 2004). But while it is true, strictly speaking, that those areas have been associated with emotion, they have also been associated with controlled cognition and with functions that seem neither emotional nor cognitive (Cera et al. 2012; Blanke et al. 2002).

2.1.1 Medial frontal gyrus

The first area Greene et al. (2001a) describe as associated with emotion is the medial frontal gyrus (MdFG; specifically, Brodmann's areas [BA] 9/10). The region does in fact seem to be involved in the processing of emotion. For instance, psychopaths show less activity than healthy subjects in the area when presented with emotionally evocative pictures (Muller et al. 2003). But the region is also "associated with high-level executive functions and decision-related processes," showing, for instance, lateralized activation in a go/no-go task (Talati and Hirsch 2005). Furthermore, it is involved in activities that seem to fall entirely outside the scope of both emotion and controlled cognition, such as the initiation and maintenance of deep sleep (Koenigs et al. 2010). In fact, the region's preferential activation[9] in 'personal' moral vignettes may not be due to its role in emotional processing *per se*, but may instead be due to its role in theory of mind (ToM) (see, e.g., Mundy 2003).

The functions of the MdFG may simply be too motley to categorize, in part because the region comprises such a large portion of the brain. In their follow-up paper, Greene et al. (2004) refer to this same region within the MdFG (BA 9/10) more specifically, as the medial prefrontal cortex (mPFC). Later, Greene (2007) focuses on two even more specific areas within the mPFC, the ventromedial prefrontal cortex (vmPFC) and frontopolar cortex (Moll et al. 2002). Such refinements in specificity are exemplary of the work Greene has been doing for the past 15 years, but with which philosophical interpretations of his work have failed to keep pace.

2.1.2 Posterior cingulate gyrus

Whereas Greene et al. (2001a) found part of the posterior cingulate gyrus (PCG; BA 31) to be preferentially activated by 'personal' moral judgments, Greene et al. (2004) later found a different (though much smaller) part of that same region (again, in BA 31) to be involved in characteristically 'impersonal' moral judgments. This is not surprising, since the PCG shows preferential activation in sundry functions (see, e.g., Hirono et al. (1998); Katayama et al. 1999; Bromm 2004).

[8] Specifically, Greene et al. (2004) claim that these prohibitions are "central to deontology," but as explained earlier (Sect. 1), this stronger claim is inaccurate and misleading; moreover, it is not relevant to the present essay.

[9] I use the term "preferential activation" to refer to the production of significantly stronger neuroimaging signals during a given task (or process; although cf. Jacoby 1991) than in a control or alternative experimental task.

While the PCG does seem to play some role in affective processes (such as the detection of threat words; Maddock and Buonocore 1997), it also appears to be involved in controlled cognitive functions, such as visual attention (Small et al. 2003) and working memory (Monks et al. 2004). At least one part of the PCG seems to be involved in emotionally valenced memory access (Vogt 2005), a function that bridges emotion and controlled cognition (Dolcos et al. 2011). This casts further doubt on the idea that the region's involvement in moral judgment can be fully explained by any dichotomous emotion/cognition dual-process model.

2.1.3 Angular gyrus[10]

Of all the areas Greene et al. (2001a) associate with emotion, perhaps the ones most strongly associated with controlled cognition are the left and right angular gyri (AG). Bilateral activation of the AG has been associated with "semantic processing, word reading and comprehension, number processing, default mode network, memory retrieval, attention and spatial cognition" (Seghier 2012). None of these functions suggests that 'personal' moral vignettes activate the region specifically because of their emotional salience, rather than their engagement of controlled cognitive processes.

The left AG appears to play a major role in mental math (Grabner et al. 2009a; Rusconi et al. 2005; Penner-Wilger and Anderson 2013; Price and Ansari 2011). It has also long been associated with reading ability (Horwitz et al. 1998), and its dysfunction has been associated with reading disorders (Pugh et al. 2000). Like the left AG, the right AG is associated with reading ability, and both its functional and physical disconnection from other brain regions have been associated with reading disorders (Horwitz et al. 1998). It has also been implicated by fMRI and transcranial magnetic stimulation (TMS) studies in mathematical thinking (Göbel et al. 2001; Grabner et al. 2009b). The region's contribution to moral processing is further clouded by its known roles in sense of agency and authorship (Farrer et al. 2008), both of which are typically thought to be important factors in both practical and normative moral deliberation.

2.2 Areas associated with controlled cognition[11]

It should be clear by now that many of the regions Greene et al. (2001a) correctly report to be associated with emotion are also involved in controlled cognition and working memory. Similarly, many regions they associate with controlled cognition appear to play important roles in emotional processing. Below, I briefly explore evidence of the affective contributions of the three areas for which Greene et al. (2001a) found greater activation in 'impersonal' moral (as well as non-moral) vignettes than in 'personal' moral vignettes.

[10] Greene et al. (2001a) list the left and right AG separately, but their patterns of activation in that paper are sufficiently similar that I discuss them together here, for the sake of brevity.

[11] Greene et al. (2001a) refer to these as "Areas associated with working memory," but in the present context it will be expedient to refer to them as associated with controlled cognition, of which working memory is one component.

2.2.1 Right middle frontal gyrus

The role of the middle frontal gyrus (MFG) in emotional and other non-cognitive processing has long been recognized. Twenty years ago, Kosslyn et al. (1996) found the region to be significantly more activated by negative emotional imagery than by neutral imagery. Greene et al. (2001a) overlook this fact when drawing on Kosslyn's work to associate certain areas with emotion (Sect. 3.2).

The MFG's role in emotional processing has more recently been implicated by work such as that of Goldin et al. (2008), Gallagher et al. (2000), and Carter et al. (2006). In particular, Carter et al.'s (2006) work suggests a role of the MFG in integrating negative emotional responses into decision-making processes, again blurring the supposed line between emotion and controlled cognition. Like the AG, the MFG seems in one sense to have been misidentified with one side of the emotion/cognition dichotomy when it much more closely aligns with the other, and seems in another sense to be altogether out of place in that dichotomy.

2.2.2 Parietal lobe[12]

Even more so than the MdFG (Sect. 2.1.1), the parietal lobe is simply too large to be strictly associated with either emotion or controlled cognition. It is usually described in terms of a dozen or so subsections. Confusingly, one of these subsections is the AG, which Greene et al. (2001a) associate with emotion (Sect. 2.1.3). To be fair though, Greene et al. (2001a) are clear that the areas of the parietal lobe they associate with working memory and controlled cognition (BA 7/40) are not within the AG. Nonetheless, it has more recently come to light that BA 7 and 40 also seem to play roles in emotional processing.

Phobic responses to both spiders (Schienle et al. 2005b) and OCD triggers (Schienle et al. 2005a) preferentially activate the right supramarginal gyrus (BA 40) within the parietal lobe, as does the perception of fearful faces (Grèzes et al. 2007). Similarly, the superior parietal lobule (BA 7) is activated to a greater degree by negative emotional film clips than by neutral clips (Goldin et al. 2008). Thus, even these areas of the parietal lobe, though they have been *associated* with working memory, cannot (strictly speaking) be *identified* with working memory.

3 Association, dissociation, and relevance in moral neuropsychology

The emotion/cognition dual-process model of neuromoral processing has enjoyed widespread acceptance in moral psychology and moral philosophy over the past 15 years. But with the benefit of hindsight, we can see that philosophers should have been more circumspect in accepting that model until more was understood about the human brain. Why, then, have philosophers so quickly and willingly come to accept this model?

[12] Greene et al. (2001a) discuss the hemispheres of the parietal lobe as two individual (left and right) regions. For brevity, I treat the entire parietal lobe as one region.

Perhaps, as I have suggested (Sect. 1), it is because contemporary moral philosophy is so deeply rooted in sentimentalism, rationalism, and the putative tension between them. Nevertheless, Greene's work really does seem to have significant implications for at least one centuries-old idea in philosophy. Ironically, it casts doubt on the idea that moral judgments based on sentiment tend to be at odds with, or even distinguishable from, moral judgments based on reason. The truth, as inelegant as it may be, is probably less stark than this.

In this section, I discuss two sorts of fallacious reasoning that may have led philosophers to prematurely embrace the emotion/cognition dual-process model of moral judgment. To keep matters simple, I continue to focus on Greene et al.'s (2001a) work in this section, reexamining the evidence and reasoning that led those authors to identify some brain regions with emotion, and others with controlled cognition.

3.1 Affirming the consequent: association, dissociation, and reverse inference

To infer the involvement of some psychological process in a task from the involvement of some previously-studied brain region in a different task thought to engage that same process is to make what is called a *reverse inference* (Poldrack 2006, 2011). Even though reverse inference is a non-formal style of reasoning (Hintikka 1998), it is essential that individuals employing reverse inference do not commit formal fallacies. Thus, a reverse inference should be discounted if it affirms the consequent. Unfortunately, many arguments for—and against—the emotion/cognition dual-process model of moral judgment are based on interpretations of neuroscientific evidence that commit this very fallacy.

Greene et al. (2001a) associate (the gyrus of) the medial frontal cortex (MdFC) with emotion, based in part on its appearance in Maddock's (1999) table of regions most frequently reported as active in emotion-related experiments in 25 publications. The MdFC ranks 12th out of 14 on that list. But as noted earlier (Sect. 2.1), the association of a given region with the processing of emotion does not imply that the region's involvement in moral processing is of an emotional nature (see, e.g., Machery 2013). To make that inference, one would at least need to dissociate activity in that region from non-emotional processing. Similarly, to infer that activity in a given brain region demonstrates the engagement of controlled cognition, one needs to dissociate activity in the region from non-cognitive processing. When interpreted with this in mind, one can see that Maddock (1999) actually undermines Greene et al.'s (2001a) claims.

Greene et al. (2001a) describe the fourth most-cited item on Maddock's (1999) list of areas associated with emotion, the MFG, as associated with working memory. And while the MFG has been associated with working memory and controlled cognition, the reverse inference that its contribution to moral processing is mediated by working memory is not justified by these associations. Rather, that claim is (more strongly) undermined by Maddock's (1999) evidence, which actually seems to suggest that the region cannot be dissociated from emotional processing.

Similarly, the regions that ranked 1st, 2nd, 3rd, 4th, and 5th on Maddock's (1999) list of regions associated with emotion were all later described by Greene and his colleagues (2004) as associated with 'impersonal' moral judgment and controlled cognition. While Greene et al.'s (2004) association of these regions with controlled

cognition is supported by the research they (2001a) cite, that same research also undermines the putative dichotomy at the heart of their dual-process model. That model requires the dissociation of those regions from emotion, a dissociation that not only Maddock (1999) but also a number of more recent papers (Sect. 2.2) undermine.

The second article Greene et al. (2001a) cite as evidence that the MdFC (BA 9) is activated specifically in response to emotional aspects of 'personal' moral vignettes—an article by Reiman et al. (1997)—is problematic in a slightly different way. Although Greene et al. (2001a) interpret the involvement of BA 9 in the recall of emotional memories *only* as evidence that the area is involved in emotional processing, Reiman et al. (1997) actually view it *also* as evidence that the region is involved in controlled cognition:

> If […] the medial prefrontal region is involved in *working memory*, it could participate in the conscious experience of emotion. If […] the medial prefrontal region participates in *executive operations*, it could participate in the initiation, monitoring, or modulation of emotion (Reiman et al. 1997; emphases added).

Thus, a rereading of Reiman et al. (1997) actually highlights just how contentious it is to assume the independence of—let alone an antagonism between—(moral) emotion on the one hand, and controlled (moral) cognition, working memory, and executive operations on the other.

3.2 Non sequiturs: relevant and irrelevant associations

In addition to affirming the consequent, researchers engaging in reverse inference are especially susceptible to non-sequiturs. Consider Greene et al.'s (2001a) interpretation of Kosslyn et al.'s (1996) finding that, compared to *imagining* negative emotional images, actually *seeing* such images elicited greater activation in the PCG.[13] Greene et al. (2001a) accurately summarize these findings by stating that the region has been "associated with emotion." But despite the accuracy of this claim, it is not all that relevant to the argument that emotion plays a key role in processing 'personal' moral vignettes. For when a person thinks about moral vignettes (of any kind), she can (but does not have to) imagine negative emotional images. However, she never actually sees such images.[14] In order to associate regional activation with emotion in the judgment of moral vignettes, one would need to find preferential activation for the recollection of *emotional* over *neutral* imagery. But Kosslyn et al. (1996) was unable to find evidence of this.[15]

[13] It is ambiguous in Greene et al. (2001a) whether the authors use Kosslyn et al. (1996) as evidence for their characterization of the PCG, or of the MdFC. Nonetheless, Kosslyn also found that the MdFC is more involved in imagining negative emotional images than in seeing them. Thus, the remainder of Sect. 3.2 applies equally well to the MdFC, if the reader replaces each instance of the work "perception" with the word "imagery," and *vice versa*.

[14] This is true, at least, for most experimental paradigms in contemporary moral (neuro)psychology, including Greene et al.'s (2001a, 2004).

[15] Ironically, Kosslyn did find that the MFG (BA 9)—which Greene et al. (2001a) associate with controlled cognition—was more activated by negative emotional as opposed to neutral imagery.

Rather, he only found evidence that the PCG is more involved in seeing negative emotional images than in imagining them. Thus, Kosslyn et al. (1996) does give us reason to believe that the region is involved in emotional processing of some stimuli, but not to think that it is involved in the emotional processing of moral vignettes. Ironically, this suggests that the PCG, despite its association with emotion, might also be involved in some 'impersonal' moral judgments, such as ones in which people are confronted with emotionally evocative visual stimuli.[16] So the blanket statement that 'personal' moral vignettes are more likely than 'impersonal' moral vignettes to elicit emotion may, if anything, be problematized by Kosslyn et al.'s (1996) findings and others like them.

4 Integrative approaches

Although I began this essay by discussing the possibility that emotional and cognitive faculties duke it out to inform moral judgment, I soon argued that the picture is more complex than that. I then discussed the mistaken reasoning that may have led so many philosophers to accept this simplistic and inaccurate model of moral processing. Chief among these reasons are the longstanding philosophical tradition of pitting emotion against reason; the helpfulness of such a model in making initial sense of empirical observations of neuromoral processing; and the infelicities of reverse inference in general. In this penultimate section of this essay, I will discuss alternative approaches to understanding the neural substrates of moral judgment, with an eye toward developing neuromoral understanding in a way that reflects these considerations, and in a way that might be more informative to philosophers.

4.1 Modeling integration and integrative models

Whatever its philosophical merits, the conception of moral judgment as a struggle to either adjudicate or find common ground between emotion and reason does not provide a good model for the empirical evidence regarding the neural underpinnings of moral judgment (Sect. 2). Because a more nuanced, sophisticated, and—I have argued—accurate view takes the perspective that moral judgments are ultimately formulated by integrating information of many kinds from many different regions and circuits in the brain, I call these integrative models of neuromoral processing. I also take such models to be integrative in a second sense, as they provide a means to integrate the most successful aspects of other models, while overcoming many of their shortcomings. In what follows, I want to consider just two examples of projects that illustrate the integrative approach I am advocating.

[16] I have in mind something like a visual representation of Greene et al.'s 'impersonal' moral vignette (2001b) in which a person blinded in war hires a doctor to kidnap a stranger and carve out his eye for transplantation. Another classic (Singer 1972) but more realistic 'impersonal' judgment might involve writing a check to help anonymous starving people; when such a decision is made after watching a commercial of dying children with distended bellies, this 'impersonal' moral judgment may be driven by emotional responses to visual stimuli.

4.2 An integrative conception of the VmPFC

Despite the criticism Joshua Greene continues to receive for his early work, he is among the first researchers whose work now reflects an integrative conception of neuromoral processing. Although his work once suggested that the involvement of the vmPFC in moral judgment reflected straightforward emotional engagement (Greene et al. 2001a, 2004), Shenhav and Greene (2014) have recently put forward a more sophisticated view of that region. They argue that the vmPFC plays an essential role in moral judgment not by underwriting aversive emotional moral judgments, but by integrating emotional information with other kinds of information shared by various other brain regions into "all things considered" (2014) judgments.

On Shenhav and Greene's view, the vmPFC is not "partisan" (2014) to social-emotional judgments; rather, the vmPFC serves an important role in integrating information that may favor such judgments with information that may favor competing judgments. The reason patients with vmPFC damage tend to make different judgments[17] about moral vignettes than healthy patients do (Ciaramelli et al. 2007; Koenigs et al. 2007), they argue, is that such patients are unable to integrate social-emotional information with other kinds of information.

4.3 An integrative conception of the rTPJ

A second example of an integrative approach is Miller et al.'s (2010) work concerning the right temporoparietal junction (rTPJ). Recent fMRI work has shown the rTPJ to be preferentially activated during judgments of individuals who act on harmful intentions but who, through sheer luck, fail to cause harm.[18] Likewise, interfering with the rTPJ via TMS has been found to reduce attitudes of condemnation toward individuals who act on harmful intentions but who fail to cause harm (Young et al. 2010). On the basis of findings like these, Miller et al. (2010) expected to find that anatomical disconnection of the rTPJ from the left hemisphere via callosotomy[19] would lead to this same unusually lenient pattern of moral judgment toward those who intend to cause harm but fail to do so. However, what they discovered was slightly more complex than they had anticipated.

Among Miller et al.'s callosotomy subjects, three had previously undergone complete callosotomies, but three only had anterior or central sections (with no immediate anatomical connection to the rTPJ) severed.[20] As predicted, patients who had under-

[17] Shenhav and Greene claim that such patients "make more utilitarian judgments" (2014), but as noted earlier (Sect. 1), I think this characterization is inaccurate and, moreover, that it is irrelevant to the present essay.

[18] A paradigmatic example of this, described by Young and Saxe (2008), is a case in which an agent attempts to put poison in someone's drink, but unbeknownst to that agent only puts sugar in the drink.

[19] A callosotomy is a procedure in which the corpus callosum, the large white matter structure connecting the left and right hemispheres of the brain, is severed. This is typically done in order to reduce the interhemispheric spread of seizure-related neural activity among sufferers of epilepsy.

[20] Depending on the region(s) in which a patient's seizures typically begin, an epileptic patient's corpus callosum may be either partially or totally severed by a surgeon.

gone total callosotomies, and hence had their rTPJ disconnected from their left hemispheres, provided much more lenient judgments than healthy participants did. But to the researchers' surprise, there was no significant difference in the judgments of the two groups of (total vs. partial) callosotomy patients in any of the experimental conditions. Both groups of patients were significantly less likely than healthy participants to condemn people who acted on bad intentions but who ended up doing no harm. Thus, it seems that a healthy connection between the rTPJ and the left hemisphere might be necessary for (the usual level of) condemnation of failed attempts to harm others, but that it is certainly not sufficient.

The most plausible explanation of these unexpected findings, Miller et al. convincingly argue, is that while the rTPJ may play a vital role in the recognition of others' mental states (and in the formation of relevant moral judgments), it must work together with more anterior regions of the brain to do so. Thus, even if the rTPJ's usual anatomical connections remain intact, its usual function will be absent unless the connections between other, more anterior regions of the brain are also intact. More recent discoveries seem to substantiate this hypothesis.

For instance, mirror neurons in the right inferior frontal gyrus (rIFG)—not the rTPJ—play an important role in the perception and modeling of others' intentions (Iacoboni et al. 2005). Incidentally, these findings may also help us recognize that Greene et al.'s (2004) characterization of the rIFG as involved in 'impersonal,' rather than 'personal' moral judgment is incomplete. Iacoboni et al. (2005) findings suggest that the region may be involved in many kinds of moral judgments, not just judgments that Greene would presumably characterize as 'impersonal.' Thus, for philosophers to characterize the region as just one more brain area driving reason as opposed to sentiment seems simplistic and mistaken.

5 Conclusion

A great deal has been written about the effort to extrapolate philosophical theories of normative ethics and metaethics from empirical discoveries regarding the neuropsychology of moral judgment. In this vein, neuroimaging discoveries have received especially strong attention from philosophers. However, I hope to have drawn attention toward deeper foundational issues concerning the extent to which these putatively value-neutral empirical discoveries have themselves shoehorned unwieldy observations into tidy but ill-fitting *a priori* philosophical frameworks. Most notably, philosophers have continued to debate the *implications* of the competitive roles of emotion and controlled cognition in moral judgment, without sufficiently questioning the supposed *fact* that their roles are neatly distinguishable—let alone competitive. A better understanding of neuropsychology as it has developed over the past 15 years suggests that both of these supposed facts may not be facts at all.

Emotion and controlled cognition are neither competitive nor clearly distinguishable in contemporary empirical findings in the neuroscience of moral processing and judgment. Therefore, attempts to understand the philosophical implications of these competitions and distinctions are fundamentally misguided. While findings within the neuroscience of moral judgment often appear to have implications for the debate

between sentimentalists and rationalists, this appearance may be due to the fact that these findings are initially interpreted according to that very sentimentalist/rationalist framework. Justifications of sentimentalism or rationalism on the basis of findings in the neuroscience of moral judgment at this time may be, to the extent that the framework of the sentimentalist/rationalist debate is superimposed *a priori* on empirical observations, largely circular.

None of this is necessarily to say that sentimentalist/rationalist debates regarding moral *philosophy* are misguided. It is only to say that the merits of such debates regarding moral *judgment* remain uncertain, and that therefore any philosophical inferences made on the basis of such claims about moral judgment should be viewed with skepticism. While the neuroscience of moral judgment may have important implications for philosophical ethics, these implications will never be fully understood until the presuppositions of philosophical ethics are no longer imposed, without satisfactory justification, on empirical observations of the brain and behavior.

For the neuroscience of moral judgment to achieve its potential in informing philosophical ethics, it must be approached in as philosophically impartial a way as possible. This means that we should embrace a variety of metaethical theories as models for moral-psychological inquiry. To do otherwise is to trivialize (i.e., to render circular) the philosophical implications of the neuroscience of moral judgment. And the first, most far-reaching presupposition with which philosophers and neuroethicists should do away is the presupposition that the debate between sentimentalists and rationalists, whatever its philosophical value, is an appropriate framework on which to hang explanations of the neuroscience of moral judgment.

Acknowledgements I am grateful to Joshua Greene and to Michael Davis, Thomas Fisher, Elisabeth Hildt, Warren Schmaus, Aaron Spink, and two anonymous referees for their comments on earlier drafts of this paper. My research was funded by a grant from the Swiss Cogito Foundation, to which I am also grateful.

References

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

Blanke, O., Ortigue, S., Landis, T., & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, *419*(6904), 269–270.

Bromm, B. (2004). The involvement of the posterior cingulate gyrus in phasic pain processing of humans. *Neuroscience Letters*, *361*(1–3), 245–249.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38.

Carter, R. M., O'Doherty, J. P., Seymour, B., Koch, C., & Dolan, R. J. (2006). Contingency awareness in human aversive conditioning involves the middle frontal gyrus. *NeuroImage*, *29*(3), 1007–1012.

Cera, N., Di Pierro, E. D., Sepede, G., Gambi, F., Perrucci, M. G., Merla, A., et al. (2012). The role of left superior parietal lobe in male sexual behavior: Dynamics of distinct components revealed by FMRI. *The Journal of Sexual Medicine*, *9*(6), 1602–1612.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264.

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

Dolcos, F., Iordan, A. D., & Dolcos, S. (2011). Neural correlates of emotion-cognition interactions: A review of evidence from brain imaging investigations. *Journal of Cognitive Psychology*, *23*(6), 669–694.

Farrer, C., Frey, S. H., Van Horn, J. D., Tunik, E., Turk, D., Inati, S., et al. (2008). The angular gyrus computes action awareness representations. *Cerebral Cortex*, *18*(2), 254–261.
Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5–15.
Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of "theory of mind" in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21.
Garrigan, B., Adlam, A. L. R., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements. *Brain and Cognition*, *108*, 88–97.
Göbel, S., Walsh, V., & Rushworth, M. F. (2001). The mental number line and the human angular gyrus. *NeuroImage*, *14*(6), 1278–1289.
Goldin, P. R., McRae, K., Ramel, W., & Gross, J. J. (2008). The neural bases of emotion regulation: Reappraisal and suppression of negative emotion. *Biological Psychiatry*, *63*(6), 577–86.
Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., Ebner, F., & Neuper, C. (2009a). To retrieve or to calculate? Left angular gyrus mediates the retrieval of arithmetic facts during problem solving. *Neuropsychologia*, *47*(2), 604–608.
Grabner, R. H., Ischebeck, A., Reishofer, G., Koschutnig, K., Delazer, M., Ebner, F., et al. (2009b). Fact learning in complex arithmetic and figural-spatial tasks: The role of the angular gyrus and its relation to mathematical competence. *Human Brain Mapping*, *30*(9), 2936–2952.
Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, *3*(3), 405–423.
Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323.
Greene, J. D. (2014a). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*(4), 695–726.
Greene, J. D. (2014b). The cognitive neuroscience of moral judgment and decision making. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences V* (pp. 1013–1024). Cambridge, MA: MIT Press.
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. (2001a). An fMRI investigation of emotional engagement in moral judgment. *Science (New York, N.Y.)*, *293*, 2105–2108.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001b). An fMRI investigation of emotional engagement in moral judgment. [Supplementary material]. *Science (New York, N.Y.)*, *293*(5537), 2105–8. http://science.sciencemag.org/content/293/5537/2105/tab-figures-data.
Grèzes, J., Pichon, S., & de Gelder, B. (2007). Perceiving fear in dynamic body expressions. *NeuroImage*, *35*(2), 959–967.
Helion, C., & Pizarro, D. A. (2015). Beyond dual-processes: the interplay of reason and emotion in moral judgment. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics* (pp. 109–125). Springer
Hintikka, J. (1998). What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles S Peirce Society*, *34*(3), 503–533.
Hirono, N., Mori, E., Ishii, K., Ikejiri, Y., Imamura, T., Shimomura, T., et al. (1998). Hypofunction in the posterior cingulate gyrus correlates with disorientation for time and place in Alzheimer's disease. *Journal of neurology, neurosurgery, and psychiatry*, *64*, 552–554.
Horwitz, B., Rumsey, J. M., & Donohue, B. C. (1998). Functional connectivity of the angular gyrus in normal reading and dyslexia. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8939–8944.
Hume, D. (2013/1751). *An enquiry concerning the principles of morals*. New York: Start Publishing LLC.
Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *Plos Biology*, *3*, e79.
Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541.
Kamm, F. M. (2009). Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs*, *37*(4), 330–345.
Kant, I. (1964/1785). Groundwork of the metaphysic of morals (HJ Paton, Trans.). *NY: Harper & Row*.
Katayama, K., Takahashi, N., Ogawara, K., & Hattori, T. (1999). Pure topographical disorientation due to right posterior cingulate lesion. *Cortex*, *35*(2), 279–282.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics*, *4*(2), 143–162.

Koenigs, M., Holliday, J., Solomon, J., & Grafman, J. (2010). Left dorsomedial frontal brain damage is associated with insomnia. *Journal of Neuroscience*, *30*(47), 16041–16043.

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

Kosslyn, S., Shin, L., & Thompson, W. (1996). Neural effects of visualizing and perceiving aversive stimuli: A PET investigation. *Neuroreport*, *7*, 1569–1576.

Machery, E. (2013). In defense of reverse inference. *The British Journal for the Philosophy of Science*, *65*(2), 251–267.

Maddock, R. J. (1999). The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. *Trends in Neurosciences*, *22*(7), 310–316.

Maddock, R. J., & Buonocore, M. H. (1997). Activation of left posterior cingulate gyrus by the auditory presentation ofthreat-related words: An fMRI study. *Psychiatry Research - Neuroimaging*, *75*(1), 1–14.

Miller, M. B., Sinnott-Armstrong, W., Young, 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., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage*, *16*(3), 696–703.

Monks, P. J., Thompson, J. M., Bullmore, E. T., Suckling, J., Brammer, M. J., Williams, S. C. R., et al. (2004). A functional MRI study of working memory task in euthymic bipolar disorder: Evidence for task-specific dysfunction. *Bipolar Disorders*, *6*(6), 550–564.

Muller, J., Sommer, M., Wagner, V., Wagner, V., Lange, K., Lange, K., et al. (2003). Abnormalities in emotion processing within cortical and subcortical regions in criminal psychopaths: Evidence from a functional magnetic resonance imaging study using pictures with emotional content. *Biological Psychiatry*, *54*(2), 152–162.

Mundy, P. (2003). Annotation: The neural basis of social impairments in autism: The role of the dorsal medial-frontal cortex and anterior cingulate system. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *44*(6), 793–809.

Penner-Wilger, M., & Anderson, M. L. (2013). The relation between finger gnosis and mathematical ability: Why redeployment of neural circuits best explains the finding. *Frontiers in Psychology*, *4*, 1–9.

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

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697.

Price, G. R., & Ansari, D. (2011). Symbol processing in the left angular gyrus: Evidence from passive perception of digits. *NeuroImage*, *57*(3), 1205–1211.

Pugh, K. R., Mencl, W. E., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Constable, R. T., et al. (2000). The angular gyrus in developmental dyslexia: Task-specific differences in functional connectivity within posterior cortex. *Psychological Science*, *11*(1), 51–56.

Reiman, E. M., Lane, R. D., Ahern, G. L., Schwartz, G. E., Davidson, R. J., Friston, K. J., et al. (1997). Neuroanatomical correlates of externally and internally generated human emotion. *American Journal of Psychiatry*, *154*(7), 918–925.

Rusconi, E., Walsh, V., & Butterworth, B. (2005). Dexterity with numbers: rTMS over left angular gyrus disrupts finger gnosis and number processing. *Neuropsychologia*, *43*(11), 1609–1624.

Schienle, A., Schäfer, A., Walter, B., Stark, R., & Vaitl, D. (2005a). Brain activation of spider phobics towards disorder-relevant, generally disgust- and fear-inducing pictures. *Neuroscience Letters*, *388*(1), 1–6.

Schienle, A., Schäfer, A., Stark, R., Walter, B., & Vaitl, D. (2005b). Gender differences in the processing of disgust- and fear-inducing pictures: An fMRI study. *Neuroreport*, *16*(3), 277–280.

Seghier, M. (2012). The angular gyrus: Multiple functions and multiple subdivisions. *The Neuroscientist*, *19*(1), 43–61.

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.

Singer, P. (1972). Famine, affluence and morality. *Philosophy and Public Affairs*, *1*(3), 229–243.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3–4), 331–352.

Small, D. M., Gitelman, D. R., Gregory, M. D., Nobre, A. C., Parrish, T. B., & Mesulam, M. M. (2003). The posterior cingulate and medial prefrontal cortex mediate the anticipatory allocation of spatial attention. *NeuroImage*, *18*(3), 633–641.

Talati, A., & Hirsch, J. (2005). Functional specialization within the medial frontal gyrus for perceptual go/no-go decisions based on "what," "when," and "where" related information: an fMRI study. *Journal of Cognitive Neuroscience*, *17*(7), 981–993.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science*, *25*(1), 111–172.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2008). Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *Journal of Neurophysioloy*, *100*, 3328–3342.

Vogt, B. A. (2005). Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience*, *6*(7), 533–544.

Young, L., Camprodon, J. A., Hauser, M., 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., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–1920.