Neuroethics manuscript No.
(will be inserted by the editor)

The Dual Track theory of Moral Decision-Making: A Critique of the Neuroimaging Evidence

Colin Klein

Received: date / Accepted: date

Abstract The dual-track theory of moral reasoning has received considerable attention due to the neuroimaging work of Greene et al. Greene et al. claimed that certain kinds of moral dilemmas activated brain regions specific to emotional responses, while others activated areas specific to cognition. This appears to indicate a dissociation between different types of moral reasoning. I re-evaluate these claims of specificity in light of subsequent empirical work. I argue that none of the cortical areas identified by Greene et al. are functionally specific: each is active in a wide variety of both cognitive and emotional tasks. I further argue that distinct activation across conditions is not strong evidence for dissociation. This undermines support for the dual-track hypothesis. I further argue that moral decision-making appears to activate a common network that underlies *self-projection*: the ability to imagine oneself from a variety of viewpoints in a variety of situations. I argue that the utilization of self-projection indicates a continuity between moral decision-making and other kinds of complex social deliberation. This may have normative consequences, but teasing them out will require careful attention to both empirical and philosophical concerns.

Keywords Morality · Neuroimaging · Reverse Inference · Self-projection

Work on this paper was supported by a fellowship from the UIC Institute for the Humanities.

Colin Klein
1420 University Hall, MC267
University of Illinois at Chicago
601 S. Morgan St.
Chicago, IL 60605
Tel.: 312-413-1801
Fax: 312-413-1093
E-mail: cvklein@uic.edu

1 Introduction

Writing of the execution of a drowsy sentinel, Adam Smith observes that:

> When the preservation of an individual is inconsistent with the safety of a multitude, nothing can be more just than that the many should be preferred to the one. Yet this punishment, how necessary soever, always appears to be excessively severe... A man of humanity must recollect himself, must make an effort, and exert his whole firmness and resolution, before he can bring himself either to inflict it, or to go along with it when it is inflicted by others. ([1] II.ii.3)

This internal conflict is familiar to philosophers. Circumstances may demand that a blameless individual suffer to prevent a larger tragedy. But even if convinced, we feel conflicted—and that conflict may even lead us to balk if it is *we* who must do the sacrificing.

Dilemmas that evoke conflicts between moral intuitions have generated a large literature. One important class of cases are the so-called *Trolley Problems*. When people are faced with the hypothetical choice of switching a runaway trolley away from a track containing five people onto a track containing only one (the classic 'Trolley Problem'), most will choose to throw the switch, killing one to save five. When faced with the choice of pushing a bystander in front of the trolley the same trolley, also saving five people, (the 'Footbridge Problem') most people say they would refuse.[1]

Why the difference? One natural explanation is that, in the Footbridge Problem, there is a conflict between

[1] See [2] for behavioral data. Selim Berker notes that the trolley problems studied in the empirical literature differ in subtle but important ways from those used in the philosophical literature ([3] 297ff), a problem that I'll pass over here.

reason and feeling. Reason declares that one must die to save many. Our sympathy for the one conflicts with this reasoned judgment; when we contemplate doing the deed in an up-close and personal way, that sympathy can be so strong as to override the judgment of reason.

The explanation just given is an example of a *dual-track theory* of moral cognition. Dual track theories typically comprise four commitments: (1) that the apparent conflict in intuitions in dilemmas is genuine: so, for example, there is no way to reconcile our intuitions in the Trolley and Footbridge cases, (2) that the conflicting intuitions issue from two distinct cognitive mechanisms ('tracks'), (3) that the cognitive mechanisms differ in systematic ways, and these systematic differences explain why they occasionally issue conflicting intuitions, and optionally (4) that at least one of the tracks is systematically unreliable, and so the intuitions issuing from it are untrustworthy bases for moral theorizing.

The story above is perhaps the simplest form of a dual track theory: moral dilemmas activate distinct cognitive and emotional tracks, and these two processes (being largely independent and concerned with different features of the world) can issue conflicting commands. This is not the only sort of dual-track model (I mention other versions in section 3), but it has proven very influential. This is in part because dual-track theories have a feature that makes them unusual among philosophical theses: commitments (2) and (3) are empirical theses, and so can be tested.

I will focus on the evidence for dual-track hypotheses that comes from functional brain imaging. This is not the only evidence for the dual-track hypothesis, but it has attracted a lot of attention since the groundbreaking work of Joshua Greene and his collaborators [2, 4]. Greene et al. investigated the dual-track hypothesis by scanning subjects as they made judgments about dilemmas like the Trolley Problem. They focused on the distinction between 'personal' and 'impersonal' dilemmas (personal dilemmas being ones where the subject had to imagine intentionally creating serious harm directly to another person), predicting that personal dilemmas would presumably be more emotionally engaging ([2] 391). They concluded that the imaging results supported the dual-track thesis. Personal moral dilemmas activated regions like the ventromedial prefrontal cortex (vmPFC), the posterior cingulate gyrus, the precuneus, and the superior temporal sulcus—all regions that had been associated with emotional processing in previous studies. Impersonal dilemmas, by contrast, created more activation in the inferior parietal lobes and the middle frontal gyrus (especially dorsolateral prefrontal cortex, or dlPFC). These are areas previously associated with working memory, a presumably more cognitive process. Hence we appear to have evidence for the dissociation that the dual-track theory would predict.

These studies have attracted intense philosophical interest, in part because Greene and others have drawn skeptical normative conclusions from the results. They argue, roughly, that some of our moral intuitions arise from the emotional track. This track is fast, sloppy, evolutionarily old, and philosophically unsophisticated. So we have reason to discount the intuitions that it issues, and with them the moral theories (like Kant's deontology) that these intuitions support [5, 6].

That conclusion is obviously controversial, and there have been a number of philosophical responses [3, 7, 8]. Most of these critiques accept Greene et al.'s characterization the neuroscience evidence, however. I believe that this is a mistake: the neuroimaging data itself is problematic, and does not support the dual-track theory.

Let me be clear about the scope of my argument. My goal is to critique the neuroimaging evidence for the dual track theory. Insofar as I attack the dual-track theory itself, I do so only indirectly, by attacking one source of evidence for it. There are other sources of evidence for the theory: lesion data, reaction-times, apparently inconsistent patterns of response to moral dilemmas, and so on. Evaluating the dual-track theory itself requires considering all of this evidence, and seeing whether, on balance, it supports the theory. Hence the present argument doesn't falsify the dual-track theory on its own: it may be that, though the neuroimaging evidence doesn't support the dual-track theory, the preponderance of evidence does.

That said, I take it as an article of faith that individual sources of evidence for a theory can and should be evaluated in isolation. As such, it is perfectly appropriate to examine whether the neuroimaging data supports the dual-track theory *without* reference to other, distinct sources of evidence. That may seem obvious: your answer to 'Why does neuroimaging data support your theory?' ought to say something besides 'Just because of all of the *other* evidence that supports my theory!' Note too that denying the independent evaluation of evidence would have disastrous consequences. Suppose that, in evaluating a source of evidence for a theory, we had to take into account all of the other evidence that had been advanced in favor of that theory. This would mean that a set of observations that jointly falsified a theory would be entirely impotent if presented serially: for each critique of one bit of evidence, one could point to the other bits of evidence to counter it. But good science is done serially, examining well-defined and limited hypotheses one at a time. So denying the independence

of sources of evidence is tantamount to asserting that most theories are unfalsifiable in practice. That's absurd. So we should hold fast to the independence of evidence.

With that in mind, I will argue for three points. First, I will argue that the empirical neuroimaging evidence—from Greene et al. and elsewhere—does not support the division of morally relevant brain areas into distinct tracks. Some of the empirical evidence that Greene et al. rely on is incomplete or outdated. Much of it has been complicated by more recent neuroimaging work. When we look at the whole picture, no dual-track theory seems to be well-supported by the empirical evidence. Second, I argue that these evidential shortcomings are due to reliance on a technique called *reverse inference*. Reverse inference uses the presence of activity due to cognitive functioning in one experimental context as an indicator of the very same cognitive functioning in a different experimental context. There is good reason to suppose that this sort of inference is illegitimate. In its place, I suggest a procedure of *cross-domain abduction*, which attributes functions to brain regions based on observed activation in the widest known variety of experimental contexts. Third, when we apply this procedure, the best explanation of the neuroimaging data is that there is a single mechanism supporting moral decision-making.[2] This unitary picture paints moral decision-making as a kind of complex self-reflexive social deliberation, consonant with other empirical and philosophical proposals.

2 Problems with Reverse Inference

2.1 Reverse Inference

I begin with a methodological point. Greene et al.'s argument crucially relies on a process known in the neuroimaging literature as 'reverse inference.' That is, they identify brain regions which are active during a particular kind of moral task, and then treat these activations as evidence for a particular cognitive theory. (Contrast this with 'forward inference,' which takes an independently confirmed cognitive theory and then uses neuroimaging data to localize functions.)

Reverse inference is controversial.[3] As many authors have noted, brain areas are *pluripotent*: that is, they can perform multiple roles and be involved in a variety of different processes. When this is the case, reverse inferences are invalid: the presence of activation is not a reliable indicator of any particular cognitive function.

It is thus crucial to Greene et al.'s argument that the brain areas they identify are not just *associated* with emotion or cognition, but *specific* to them. That is, they must activate if *and only if* a subject is engaged in a characteristic emotional/cognitive process. Conversely, if these regions are not specific, then the neuroimaging evidence is an extremely weak test of the dual-track hypothesis.

Given the pluripotency of many brain areas, claims of specificity should be taken with a grain of salt. Further, such claims are hostage to empirical fortune: new discoveries can reveal additional complexity in an area once thought to be specific. I argue that fortune has not been kind to the dual-track hypothesis. Recent empirical evidence shows that none of the areas identified by Greene et al. are specific to either emotion or cognition.

2.2 Parietal Areas

2.2.1 Precuneus/PCC

Greene et al. found increased activation in the Precuneus (BA 7, hereafter PC) and the posterior part of the cingulate cortex (BA 23 and 32, hereafter PCC) during personal moral dilemmas. They argue that this represents activation of emotional areas.[4] The assignment of PC and PCC to emotional processing, however, is done on slim grounds.[5] There has been considerable work on PC and PCC in the past decade, and the claim that PC/PCC makes for a specifically emotional processing area is no longer sustainable.

First, the PC. Kober et al.'s recent, wide-ranging meta-analysis of emotional processing did not include

[2] I will use 'decision-making' rather than the more common 'deliberation' because 'deliberation' is ambiguous between reflection on particular cases and reflection on general rules. The studies I discuss have focused on decisions about individual cases. This may involve reflection on general principles, but need not.

[3] For critiques, see [9–11]. The 'inference' terminology needn't be taken literally: one can do reverse inference probabilistically by showing that some cognitive process was *likely* to have occurred given some brain activation. There is good empirical evidence that probabilistic reverse inference is equally problematic [12].

[4] This causes their own theory some grief—PC activation was also found to be most active during utilitarian moral judgments. They argue that perhaps all action requires some affective motivation, and that these areas provide this in the case of utilitarian judgment ([2] 397). They provide no independent empirical evidence for this claim, though. Further, as Berker notes, establishing this would be of dubious help to their theory as "what is at stake here is whether *all moral judgment*, not *all action*, has an affective basis." ([3] 307).

[5] Greene et al. cite only Maddock's meta-review of emotion-processing studies in support [13]. However, Maddock specifically excludes the PC from his argument: his focus is on retrosplenial cortex, and he uses the inferior edge of the precuneus as one of his limiting boundaries ([13] 310). Further, Maddock's argument that PCC is *specific* to emotion processing is based only on the fact that he couldn't find any non-emotion studies that activated it ([13] 313). That argument is, as I will show, no longer sustainable.

the PC among emotion-associated networks (See [14] Fig 7). Along with this lack of evidence for emotion-specificity, there is considerable positive evidence about the functions that the PC implements. Cavanna and Trimble did a meta-review of studies with PC activation and found that PC was most commonly associated with three types of task: visual imagery, successful episodic memory retrieval (both visual and non-visual), and self-referential processing, especially when that involved an experience of agency [15]. Self-referential processing in the PC has in turn been associated with general awareness and consciousness; indeed, PC has one of the highest resting metabolic rates of any portion of the brain [16]. This self-referential processing is also strongly implicated in taking a first-person perspective during tasks [17, 18]. Taking a first-person perspective may also play an important role in theory of mind, or understanding others' thoughts and intentions. The PC has been directly implicated in theory of mind tasks, and in taking the perspective of others when viewing painful scenarios [19, 20]. It has also been shown to be active during attribution of emotion to both self and others, which Ochsner et al. explain by the use of the first-person perspective to successfully understand others' experiences [21].

The connection between episodic memory retrieval and self-referential processing may also be more than coincidental. The PC is active when subjects think about both the future and the past [22], and Addis et al. found that the PC was most active when subjects mentally added details to imagined past (or future) events [23]. Johnson et al. found that the PC was active during self-reflective processing both about one's goals and about the duties that constrain one's activities [24]. Thus, the PC may utilize episodic memories in order to flesh out details of self- and other-oriented cognition, including more classically cognitive reflections on duties, goals, and so on.

The situation with the PCC mirrors that of the PC in many ways. Kober et al.'s meta-analysis of emotion networks did include the PCC. However, they provide a number of reasons to be wary of labeling the PCC as a specifically emotional area. Functionally speaking, PCC seems to play a nonspecific relay role between other emotional-related networks, instead subserving a number of attentional and perspective-taking functions ([14] 1014).

Many functions of PCC appear to be complimentary to those of the PC. In his meta-review, Maddock suggested that PCC activation during emotional tasks might represent the recall of episodic memories in order to evaluate emotionally salient events ([13] 315). PCC activation is seen alongside PC activation in elaboration of first-person stories [23]. It is also seen when subjects must attribute emotion to self and others [21]. It is also activated in making judgments about whether character traits apply to oneself or to others [25]. Like the PC, it also has a high resting metabolic rate, and seems to be a part of 'default mode' processing that is a prerequisite for conscious awareness [16, 26, 27]. A recent review of PCC function thus concluded that it plays a variety of cognitive roles similar to that of the PC, including underlying imagination and memory, navigation, scene transformation, and so on [28].

There is thus little evidence that the PC or the PCC plays an emotion-specific role.[6] They both may have something to do with emotional experience: it would be a surprise if self-reflexive processing and imagination wasn't related to emotion in some broad sense. However, the *type* of processing that both the PC and PCC engage in clearly involves sophisticated representations of self, others, and the world. As such, identifying them as purely emotional areas is a stretch: they perform functions that underly a variety of different cognitive processes.

2.2.2 Superior Temporal Sulcus

The identification of the Superior Temporal Sulcus (STS)[7] as a specifically emotional area is also implausible. STS plays a number of well-known roles in the low-level evaluation of socially salient stimuli. The activation in Greene et al. is relatively posterior and dorsal; this portion of the STS is most commonly associated with audio-visual integration, theory of mind tasks, and the processing of faces, with some role in processing biological motion as well [30]. It appears to be especially important for processing socially communicative clues like eye gaze [31] and finger-pointing [32]. Abnormal activity in the STS is hypothesized to be partially responsible for the social abnormalities in autism [33].[8] Thus, there is little evidence that STS activation is specific

[6] Later fMRI investigations of moral decision-making have come to a similar conclusion; see for example ([29] 812).

[7] Greene et al. identify the STS in their 2004 work. A similar region with the same Brodmann's area, though with a slightly more dorsal extent, is identified as the Angular Gyrus in the 2001 study. While Greene et al. did not provide activation foci for the angular gyrus in their 2001 paper, they consider the 2004 STS activation (the only 'personal' activation in BA 39) a replication of the results of their 2001 study ([4] 391). I shall follow them in doing so, and so identify the activation in the two areas.

[8] The two studies that Greene et al. cite in favor of STS being an area specialized for aversive stimuli both used visual presentations of people [34, 35]; the activation attributed to emotion in these experiments can also be explained by enhanced attention to salient facts about human figures.

to emotional processing in a way that would support a dual-track theory.[9]

2.2.3 Inferior Parietal Lobe

In both studies, Greene et al. found activation in the inferior parietal lobe (IPL) that was more active during impersonal moral decision-making. They attribute this activation to cognitive processing, and in particular to working memory activation. There is evidence that the IPL supports working memory function [36]. However, recent evidence now suggests that IPL function is considerably more complex.

On the one hand, IPL activation also appears in emotionally laden moral judgment tasks. Borg et al. found inferior parietal activation when subjects considered sociomoral transgressions. They also found a focus in the IPL—very close to that reported by Greene et al.—that was most active during consideration of incest scenarios [37]. Incest scenarios are a stereotypical example of moral judgments that also evoke strong emotional reactions [38]. As such, even if we accept a distinction between emotion and cognition, it appears that the IPL plays a role in both.

On the other hand, activation foci that Greene et al. found lie, along with the STS, in a region known as the temporo-parietal junction (TPJ). As with the STS, there is now considerable evidence that the inferior parietal portion of the TPJ is crucial for social reasoning, and particularly for attributing mental states to others. It is consistently activated by tasks that require attributing beliefs to other agents [20, 39].[10] Patients with lesions in the left TPJ are profoundly impaired when reasoning about others' false beliefs [40]. And, crucially, experiments by Young and Saxe have shown that IPL is most active during the 'encoding' phase of moral decision-making—that is, when subjects have to form beliefs about the intentions of the agents involved in particular scenarios [41, 42].

These observations cast doubt on Greene et al.'s attribution in three ways. First, even if IPL is sometimes involved in working memory, it is not specifically engaged for that task. Hence claims of psychophysical invariance fail. Second, there is another plausible interpretation of IPL activity: that it is involved (with STS) in the social cognition necessary to make sense of moral scenarios. Third, this social evaluation function needn't be specifically either 'emotional' *or* 'cognitive': it could well underly emotional responses based on others' mental states as well as more dispassionate reasoning about others' intentions.

2.3 Frontal Areas

2.3.1 Medial PFC

Greene et al. identify the ventromedial area of the PFC as another emotion area. This connects their work to well-known work by Damasio and Bechara, who have argued that vmPFC plays a crucial role in emotion monitoring and the use of emotions to guide choice [43]. Consonant with Greene's hypothesis is some evidence that patients with vmPFC damage make more 'utilitarian' choices during moral reasoning tasks, though the interpretation of these results is hotly debated [44–48].

Identifying the vmPFC as a *specifically* emotional area, however, is problematic. For one, vmPFC patients are also profoundly impaired when it comes to planning and executing actions, which suggests that the vmPFC is also essential to successful cognition and deliberation [49].[11]

Further, there is increasing evidence that frontal cortex does not have sharp functional divisions, but is rather organized along gradients of more or less abstract representations and process-types, with regions near the frontal pole concerned with the most abstract sorts of representations [50, 51]. Amodio and Frith argue that a similar division is apparent in medial PFC.

[9] It is worth highlighting a subtle but important shift in the description of STS in Greene et al.'s work between 2001 and 2004. The areas associated with personal moral dilemmas are, in 2001, attributed to emotional processing in an unqualified sense. In 2004, however, they are identified as areas important to "emotion and social cognition" ([2] 391). One might think that the distinction between the two is rather important. 'Social cognition,' at least as performed in the STS, can involve straightforwardly cognitive processes—that is, it may just as well support rarefied deliberation about other's goals as well as more immediate, emotionally laden interactions . Further, it is one thing to set up 'emotions' against proper moral decision-making—that at least has a long history. But it's not obvious why or how ordinary moral deliberation could be distinct from social cognition: on nearly every theory of moral cognition (including the utilitarian), moral decision-making centrally involves thinking about our obligations to others to whom we are socially related (I will return to this theme below).

[10] This activation is relative to similarly difficult tasks that require attributing false representations to nonintentional objects like pictures, suggesting that working memory effects are unlikely to explain the increase.

[11] The deficits in vmPFC patients are also problematic for Greene's normative project. Berker puts the point well when he notes that: "it is dialectically problematic first to appeal to patients with damage to emotional brain regions when making an empirical case for the dual-process hypothesis and then to go on to argue that the verdicts of these brain regions should be neglected (in effect urging us to be more like these patients), since many patients with this sort of brain damage make moral decisions in their personal lives that count as disastrous when evaluated by just about any plausible normative standard." ([3] 314).

More dorsal areas represent the value of future actions, while more ventral regions represent the value of outcomes ([52] 270). Within the vmPFC, there is evidence that different evaluations are represented via a 'common currency': that is, valuations of outcomes from a variety of sources (deliberation, emotional evaluation, or whatever) are represented in a commensurate way [53, 54]. This means that activation in the vmPFC and frontal poles need not reveal anything about its source, but only about its role in future deliberation.

Further, while some moral reasoning studies do find activation in the really ventral parts of lateral PFC (for example [55] and [56]), the activation foci in Greene et al. are actually more towards the medial-lateral PFC, in an area that Amodio and Frith dub the anterior rostral PFC (arPFC).[12] Activation in this more dorsal area is seen in a number of other imaging studies of moral cognition [37, 41, 42, 57, 58]. Amodio and Frith note that the arPFC is activated in studies that require self-report of emotion, but note that "Such commonly-used 'emotion' tasks overlap significantly with tasks assessing self-knowledge—that is, being asked to report one's emotional response is essentially a question about self-knowledge" ([52] 272). They also note that arPFC is active in many studies that do not involve strong emotional responses, and argue that the characterization of the arPFC as an emotion region is therefore "not appropriate" (272).

Supporting this assertion, the arPFC appears to have a different functional profile from more ventral and dorsal regions of mPFC, and plays an especially important role in reflection on the mental states of oneself and others. Ochsner et al. found increased activation in arPFC when subjects had to either self-attribute emotional responses to a scene or to attribute emotional responses to a person in a scene (compared to judgments of the scene's location) [21]. A similar pattern was shown when subjects had to make attributions of adjectives to themselves or to socially close others [25]. Oschner et al. performed a large meta-review of studies that found dorsomedial PFC activation and discovered that a large number of them found arPFC activation for studies that involved meta-reflection on the mental states of self or other. They conclude that arPFC activation "might be important for the metacognitive ability to re-represent affective, cognitive, and other types of inputs in a self-generated symbolic... format" ([21] 9). Vogeley and Fink in turn argue that this ability is one of the reasons why medial PFC is strongly implicated in first-person perspective-taking [17].

[12] Amodio and Frith set the boundaries of arPFC at Talairach z=2 to about z=45 ([52] 270). Greene et al.'s foci are at z=17 and z=19 in 2001 and 2004, respectively.

This self-reflective cognition also appears to be important for social theorizing. Activation in arPFC is found when subjects perform theory of mind tasks [20, 41]. Mitchell et al. found that this effect was strongest when subjects mentalized about socially close others, suggesting that the attribution of theory of mind involved self-reflection on one's own mental states as a paradigm [59]. Medial PFC also appears to be selective for self-evaluations that require complex thought about others' mental states. Moll et al. found arPFC-specific action for imagined embarrassment, guilt, and compassion. They label these the 'pro-social' emotions, on the plausible hypothesis that feeling them crucially requires taking a stand on others' mental states. Similarly, Robertson et al. found arPFC activation associated with detection of moral issues of care or justice in business ethics stories [60]. Finger et al.'s results build on this picture further; they found arPFC activation when subjects imagined performing moral transgressions or social transgressions with an audience present; they suggest that this activation might represent determining the correct response to situations where behavior towards others must be changed [61].

Two themes should be apparent. First, mPFC activation subserves a wide variety of self-reflective cognitive functions, not just emotional ones. Second, insofar as mPFC is associated with emotional processing, it can be associated with complex and subtle higher-order emotional processing: that is, in forming appraisals and beliefs about emotions, not just in feeling them. Medial prefrontal cortex is thus not an area exclusively concerned with simple, non-cognitive emotional processing.

2.3.2 Dorsolateral PFC

Greene et al. identify dlPFC (BA 46) as an unambiguously cognitive area based on its role in working memory. Meta-analyses of working memory do identify dlPFC as an area crucial to working memory [36]. However, we must distinguish between two hypothesis about the function of dlPFC during moral decision-making. On the one hand, dlPFC could be a *source* of moral intuitions. This appears to be the interpretation of Greene et al. in their 2001 paper; having identified dlPFC as a cognitive area, they use this as evidence that impersonal moral dilemmas stem from cognitive deliberations. But dlPFC (along with the ACC) also plays a crucial role in executive function, and in particular in maintaining representations of conflicting demands for action.

I argue that there is good evidence that, at least in moral decision-making, dlPFC plays latter, adjudicatory role. That is, there is evidence that dlPFC in

not part of 'core' moral reasoning but rather provides working memory resources that are recruited to aid core moral reasoning when necessary. First, Greene et al. note that dlPFC is preferentially active along with the ACC during difficult (rather than easy) personal moral dilemmas, precisely as one might expect for a general adjudicatory area. Prehn et al. found that activation in dlPFC inversely correlates with 'moral competence,' a measure of an individual's ability to appreciate different abstract moral considerations [56]. Similarly, Heekeren found a correlation with reaction time in dlPFC in moral decision-making tasks that was independent of task type or presence of bodily harm [55]. Both results suggest that dlPFC works harder the harder an agent must work to resolve a problem. There is also evidence that sociopaths, who appear to be impaired in ordinary moral reasoning, show increased activity in lateral PFC, possibly as a compensation mechanism.[13] Given this, one should expect dlPFC to be more active in situations where conflict is more salient—arguably, precisely the conditions that obtain in Greene et al.'s 'personal' dilemmas.

If dlPFC plays an adjudicatory role, then dlPFC activation is only evidence for the dual track thesis if it reliably signals a conflict between emotion and cognition. Unfortunately, it does not. dlPFC activation is also seen when there is conflict between stimuli with competing emotional valence [64, 65] and between different abstract rules [66, 67]. That is, dlPFC activity is seen when the conflict is between *entirely* emotional or *entirely* cognitive considerations. Thus dlPFC activity is not obviously either cognitive or emotional: it can adjudicate between actions that have a wide variety of sources, and is modulated by both cognitive and emotional inputs ([68] 105ff). So while dlPFC activation may indicate a conflict to be resolved, it does not indicate the *nature* of that conflict. It may be between the product of two distinct processes, or between conflicting contents of the same system (in the way, for example, the conflict caused by grammatically ambiguous sentences is a conflict within a single system for syntactic parsing). It is possible, for example, that the conflict in personal moral dilemmas is nothing more than a conflict between two distinct rules ('Don't put people in danger' versus 'Save many even at the cost of a few') that are themselves the product of a unified system.

Thus, there is good reason to be suspicious of the identification between dlPFC and a specific kind of moral deliberation. Imaging studies of moral reasoning have used stimuli ranging from the simple to the tricky; observed dlPFC activation may simply represent the recruitment of a general resource to compensate for unfamiliar, difficult scenarios.[14] Given this, dlPFC activity provides neither direct nor indirect evidence for a dual-track theory.

[13] See the exchange between [62] and [63] for a discussion.

2.4 Limbic Areas

A final word is necessary about limbic activation. Studies of moral reasoning occasionally report activation in limbic and paralimbic areas. These structures—most notably the amygdala, the insula and the paracingulate cortex—are often associated with emotions like fear, disgust, and so on. These activations in particular have captured the attention of philosophers: Woodward and Allman's discussion of empirical work in moral decision-making, for example, focuses on the activation in these areas during personal moral deliberation almost to the exclusion of the other cortical activations that Greene et al. demonstrated.[15] However, the significance of this activation is harder to interpret than one might suppose.

First, many of the studies that report limbic activity use stimuli deliberately designed to provoke strong emotions like anger, disgust, and so on. There should be little doubt that the stimuli are effective: subjects do have the intended reactions. But this means that at least some emotion-related activation should be expected *regardless* of the functional hypotheses you hold. This activation does not favor any particular functional (or philosophical) hypothesis [70]. Both a dual-track theory and its opponents should expect (say) insular activity in response to moral scenarios that evoke disgust, just because the insula seems to be especially sensitive to these sorts of stimuli. Limbic areas seem to be among those reliably activated when subjects passively view morally relevant stimuli without the need for judgment [57], again suggesting that their role is primarily in detecting salient features of situations. To put it

[14] This also allows an alternative explanation of the reaction time data in [69]. Greene et al. observe that cognitive load selectively slows utilitarian judgments. However, scenarios in which a purely utilitarian choice is the plausible one are, arguably, relatively unfamiliar in everyday life. One should expect a compensatory mechanism to be more active in these cases, and reaction times to be correspondingly slower when under load. Greene et al.'s alternative interpretation, that working memory is required for utilitarian judgment *per se*, seems implausible for two reasons. First, increasing cognitive load only increased reaction time, not the proportion of non-utilitarian judgments, contrary to what one would expect from interference with a functionally crucial subcomponent. Second, the subpopulation of subjects who gave the most utilitarian judgments actually responded faster than those who gave non-utilitarian judgments, contrary to what one would expect if utilitarian judgment essentially required slow working memory processes.

[15] For an example, see ([7] 167).

another way: activation in limbic areas to emotionally laden stimuli can be entirely explained by differences in the *stimuli*, not in the processes that these stimuli provoke.

Second, limbic activation isn't necessarily specific to emotional response. Consider, for example, the insula. The insula appears to be involved in a complex re-representation of the body in order to support interoception about bodily and emotional states in a way that promotes action [71, 72]. Sanfey et al. discovered that the insula is also active when subjects reason about complex economic games [73]. Reasoning about unfair offers in the ultimatum game requires deciding whether or not to reject an 'unfair' offer that would nevertheless bring one some money; subjects typically choose to reject such offers. Behavioral evidence suggests that subjects reject unfair offers in order to preserve both their reputation and pride [74, 75]. This sort of reaction is, arguably, more sophisticated and cognitive than reactions like disgust—for starters, it requires higher-order representations of others' beliefs, desires, and so on. Part of the insula's function presumably involves representation about bodily integrity: some patients with insular damage can, notably, recognize threats like painful stimuli but are unmotivated to do anything about them [76]. Feelings of pride may in turn involve notions of bodily integrity in an extended, metaphorical sense. Recent work similarly suggests that the amygdala is involved in a variety of processes, including cognitive ones like attentional modulation and salience-marking. ([68] 148ff; [77] 186ff).

The presence of limbic and paralimbic activation is thus not an invariant marker for the absence of cognition [78]. Such activation may provoke and support cognitive attitudes like pride and reputation-maintenance. At the very least, the possibility that they may do so undercuts the *normative* significance of these activations.

Thus, a simple interpretation of limbic activity is uninformative; a complex interpretation invites a complex philosophical response. For what it's worth, Greene et al. put relatively little weight on these activations, and I think for exactly the right reason. They note that the insula tended to be more active when subjects contemplate difficult personal moral dilemmas, but they suggest that this probably has to do with the increased time spent considering repugnant acts. As the insula "subserves negative affective states," increased activation would be expected ([2] 395). Greene et al. spend considerably more time discussing the significance of activation in the cerebral cortices, as these are more plausibly the ones that underly moral *decision-making*, rather than simple reactions to features of moral situations. So if the case for a dual-track theory is to be made, it must be made via other cortical areas. As I have shown, that case cannot be made.

3 Neuroimaging and Taxonomy

I have argued that none of the areas identified by Greene et al. are specific to either emotion or cognition. As such, reverse inference from neuroimaging data to a emotion/cognition dual track theory is illegitimate. In some sense, this should not be a surprise. Functional imaging evidence increasingly suggests that few, if any, brain areas are specific to either cognition or emotional responses [68, 79]. Instead, there is evidence that every area of the brain is modulated by emotional state, and every area of the brain can be affected by classically 'cognitive' processes.

Greene et al. may appear to be on solid ground in a more general sense, however. For personal and impersonal moral judgments did appear to activate *distinct* networks, whatever the best description of those networks might be. Part of the allure of fMRI, as Camerer et al. put it, is that it allows us to look inside the 'black box' of the brain ([80] 9). This may in turn provide evidence about the number and organization of distinct processes, independent of what the function of these processes might be. Call these hypotheses about the number and classification of mental processes *taxonomic* hypotheses, as distinct from *functional* hypotheses about the causal relationships between areas so identified. If we really do have evidence for the taxonomic hypothesis of distinctness, some of Greene's positive normative project might remain plausible.

Yet while the imaging evidence may be *consistent* with a dual-track model, we would be too hasty to conclude that it *supports* a dual-track theory. Consistency is an extremely weak standard against which to evaluate neuroimaging evidence. Instead, Greene et al. would need to show that the imaging evidence *favors* a dual-track model over its rivals [70, 81]). Evidence can be consistent with a hypothesis even as it makes it less likely; only when we consider a hypothesis against its rivals do we know whether it is supported by imaging evidence.

One salient alternative to the dual-track model is a single-track model—for example, that suggested by Jorge Moll and his colleagues [82, 83]. On a single-track model, moral decision-making is implemented by a single, unified set of brain areas. Different portions of this set may be responsible for different aspects of moral

cognition, including the evaluation of different kinds of reasons.[16]

Importantly, these distinct reasons are commensurate: they can be based on the same kinds of evidence, evaluated against one another, and integrated together into complex actions. Contrast this with the dual-track view, on which the activity of the two tracks is incommensurate and often opposed. On the single-track view moral decision-making is the process of integrating a set of different considerations; on the dual-track view, it can only be making a decision about which track will win out.

The mere presence of differential activation during different moral tasks is entirely consistent with—indeed, predicted by—a single-track hypothesis. Evidence that a brain area R shows greater activation in A than B is ambiguous. It could mean that R is inactive in B, or it could just mean that R is active—and functionally important—during B, but less so than during A. Single-track theories predict that different portions of a common network will be more or less active depending on the details of the stimulus: impersonal moral dilemmas, say, might cause more activation in areas that require shifting to a third-person perspective. But such activation will, on a single-track theory, also be important for evaluating personal moral dilemmas—just less important than it is for evaluating impersonal ones.[17].

The only way that differences in activation could show functional distinctness would be if we had good reason to suppose that A and B should produce *exactly* the same amount of activity in the two conditions: that is, that the null hypothesis for a single-track theory is identical activation across conditions. But this is a dubious assumption, and one denied by extant single-track theories.[18] On the single track view, differences in activation simply reflect differences in what the stimuli demand. As such, we should expect imaging experiments to fractionate the single network in various ways depending on the differences in the stimuli. Evidence of fractionation thus does not itself weigh against the single-track view.

Indeed, if our theorizing focuses only on differential activation in particular brain areas, it is unclear whether *any* cognitive theory is favored over an indefinite number of possible rivals. As I noted above, most brain areas are pluripotent: there is a many-to-one mapping between cognitive functions and brain areas. If a task activates even a small number of brain areas, this leads to a combinatorial explosion of possible cognitive theories, one for each possible permutation of functional attributions. Many of these combinations will represent implausible cognitive theories. But the sheer number of possible interpretations weaken our confidence that any particular cognitive theory is favored by the data. Further, mere consistency of imaging data with a cognitive hypothesis no longer represents a rigorous test of that hypothesis: there will always be too many different ways to re-interpret the data to fit what is observed.

In earlier work, I suggested that this was a reason to be skeptical about difference-based fMRI analyses (though not about more sophisticated methods of data analysis) [87]. I now think that may be hasty. Reverse inference can be seen as a form of inference to the best explanation (IBE): it claims that the best explanation of activation seen in some experimental context B is that the brain region was performing exactly the same function as in some earlier context A.[19] As a form of IBE, it is fundamentally flawed: it does not take into account other functions that a pluripotent region might be performing, and therefore cannot claim to be the *best* explanation of observed activation.

The solution, I argue, is to move to what I'll call *cross-domain abduction*. In this, I follow the lead of Price and Friston's analysis of posterior lateral fusiform in their excellent [11]. Price and Friston note that PLF is active in a dizzying variety of tasks: viewing words, picture naming, making unprimed semantic decisions, decoding braille, and so on. Taking PLF activity as indicating the presence of any one of these cognitive tasks would be problematic, for precisely the reasons I've indicated. However, one can use this data to argue that PLF performs a more general function—what Price and Friston term sensorimotor integration—that underlies a variety of superficially distinct cognitive processes. Note that this strategy differs in important ways from reverse inference. Reverse inference takes activation in a *single* experimental context, and uses it (illegitimately) as an indicator for the presence of a particular cognitive operation. By contrast, the present strategy takes activation

[16] Moll and de Oliveira-Souza, for example, suggest that vmPFC might be more responsible for the evaluation of other-regarding prosocial reasons for action, while the vlPFC evaluates other-critical reasons responsible for resentment and anger [48].

[17] This is a general problem facing the direct test of dual-track theories. Similar critiques have been posed for dual-track theories of memory ([84]) and first- and second-language acquisition ([85, 85]

[18] This reduces to a more general concern about the use of null hypothesis significance testing; see [86] for the general concern, and [87, 88] for discussions in the context of neuroimaging.

[19] Note that this would avoid the problem with deductive readings of reverse inference, namely that they appear to be straightforwardly invalid because they affirm the consequent ('If function F is performed, then A is active; A is active, therefore F was performed') [12]. I commit to nothing further about how IBE itself should be understood; I'm inclined to think that it will itself be cashed out in probabilistic terms, but that is irrelevant for the present purpose.

across a *distinct* set of contexts, and presents a theory that provides the best explanation of that activation across all contexts. As a species of inference to the best explanation, of course, the evidence is not knockdown. By incorporating activity across the widest possible set of experimental contexts, however, we can be more confident in our attribution of function to a brain region, and thereby avoid the obvious problems with reverse inference.

Note that cross-domain abduction crucially requires drawing in observations made in a wide variety of experimental tasks. It is insufficient to appeal to a mass of evidence, however large, drawn from related tasks. The *number* of experiments which attribute a specific function to a brain region depends only on the experiments that scientists have found interesting to do. In cross-domain abduction, they collectively count for no more than single experiments. So (for example), suppose that we found that the PPC was active in dozens experiments involving emotional response. A cross-domain abduction need not count these more than a single experiment, The explanatory target is the fact that PCC activation appears in emotional tasks, *and* in retrieval of episodic memory, *and* when elaborating first-person stories, and so on.

In the case of complex cognitive tasks like moral decision-making, cross-domain abduction must be applied to *sets* of commonly activated regions, rather than a single region. The same logic applies, however: the goal is to explain why the same set of regions is activated across a variety of superficially distinct experimental contexts. Further, if we take the set of activated regions to form a network, we can tease out the contributions of individual areas after we have attributed a general function to the network as a whole. Brain regions are functionally pluripotent in part because individual brain areas dynamically combine to form networks that jointly perform cognitive tasks ([68, 89]). By starting with the function of the network as a whole, we can work back to give plausible general explanations of the function of each region in the network. Again, these functional attributions will avoid the straightforward problems with reverse inference, precisely because in forming them we incorporate observed activation across the widest possible set of experiments.

That was all a bit abstract. In this section, I will flesh it out by arguing that a single-track interpretation of the data is best supported by cross-domain abduction.

4 A Positive Story: Moral Decision-Making as Self-projection

4.1 The Core Self-Projection Network

There is increasing evidence that a single brain network supports what Buckner and Carroll call *self-projection*: a shift of perspective to an alternative, non-actual environment that is referenced to oneself [90]. This core network for self-projection comprises the posterior cingulate cortex and precuneus, the inferior parietal lobes, the temporo-parietal junction, the medial prefrontal cortex and the more rostral portions of the orbitofrontal cortex, and the medial temporal lobes. This network is active during a number of cognitive tasks. These include remembering the past, imagining the future and planning for action, navigating in complex environments, and theorizing about the mental states of others, especially in socially complex situations. Each of these four types of task reliably activates the core brain network when subjects are scanned [90, 91].

It should not be terribly surprising that these tasks share a common pathway. It has long been theorized, for example, that imagination and episodic memory are aspects of a unified creative process [92]. Supporting this view, amnesiacs are often impaired not just in episodic memory, but also in their ability to imagine themselves in the future, suggesting a common substrate for both functions [93].[20]

Further, there is a theoretical similarity between these tasks. Each requires, in a broad sense, representing the world from perspective other than the one that you actually occupy. Memory requires representing the past from the perspective of a past self. Action planning requires imagining potential future worlds from the perspective of one's future self. Navigation requires representing the actual world from a perspective that one does not currently occupy. Reasoning about complex social situations often requires representing the world from the perspective of a distinct agent, and imagining how their beliefs and actions might depend on what you do. In each case, there is a common structure. One must (1) generate and maintain a representation of the world as from a particular point of view, (2) maintain this representation even though it is different from the way things actually are, and (3) appreciate the relationship between one's current state and the represented state, so that one might evaluate or otherwise cognize about the represented state. Self-projection thus appears to

[20] Schacter et al. discuss similar associations [91]. Also interesting for the present discussion is [94], which discusses a case in which a subject had both amnesia and failed self-regulation in social situations.

be a determinable type of cognitive function, of which particular applications like episodic memory retrieval are determinate instances. The differences between determinate applications of self-projection involve differences in the type of representation and the perspective on it.

Consonant with the argument in section 3, different self-projection tasks also differentially engage parts of the common network. So, for example, Addis, Wong, and Schacter found the posterior portions of the core network were preferentially activated during event construction while frontal networks played a larger role in event elaboration [23]. Okuda et al similarly found that the frontal portions of the network were more strongly engaged by future-oriented deliberation [22], and Ochsner et al. argue that the medial PFC is especially strongly engaged in prospective social reasoning [25].

This sort of empirical evidence provides a corroboration of the cross-domain abduction. Focusing on theoretically similar self-projection tasks allows us to identify a core brain network, the activation of which distinguishes them from other, potentially similar tasks. By looking at differential activation within that core network, we can build a theory about how subregions of the network contribute to the overall task. Differential activation across subtasks is not, however, evidence for multiple distinct processes: instead, it shows only that different kinds of self-projection require more or less engagement of core resources.

4.2 Self-Projection and Moral Deliberation

The core areas involved in self-projection are also the areas that are found as parts of the network recruited for moral cognition. Table 1 shows that these areas are specifically active in a wide variety of moral tasks. Not every area is active in each study, but that's to be expected: differences in statistical power and study contrast make unanimous agreement unlikely. Nevertheless, the frequent conjoint activation of these four groups of areas is suggestive of an underlying unified network.

Young and Saxe recently looked at this core network in the context of moral judgment. They found that consideration of complex moral scenarios recruited the network for self-projection [41]. Further, they found two broad distinctions within these networks. The more posterior bits of the network were preferentially recruited during an early, encoding phase: that is, when subjects first formed a mental picture of a presented scenario. Portions of medial PFC became more active in the later phases, when subjects had to integrate knowledge about the situation to form a fully elaborated scenario that could then be judged. Harenski et al. come to a similar conclusion in recent work, noting that posterior brain regions are engaged even in implicit detection of morally salient features, while medial frontal networks appear preferential activated by explicit moral deliberation [95].This fits well with a view on which the TPJ, STS, and precuneus play an important role in understanding scenarios that require drawing on a theory of mind [20], while the frontal cortex plays a role in integrating this knowledge with the subject's goals and social commitments [83]. Recent work by Young et al. has offered further evidence for this: disruption of the TPJ using TMS during moral decision-making tasks greatly reduces a subject's ability to take into account an agents beliefs and intentions when making moral decisions [96].

Rather than two distinct networks involved in moral cognition, experiments on moral deliberation reveal portions of a single, unified network that is involved in prospective social cognition. Again, it is important to emphasize that the evidence for a single network is entirely compatible with the fact that only some portions of the network are seen activated in various experiments. Even if the same network is involved in all social cognition, it is exceedingly unlikely that every area will be involved to the same degree on every task—some tasks will require more imagination, others more integration, others greater sensitivity to others' mental states, and so on. This does not preclude experiments that try to tease out the relative contributions of different areas within the core network by using stimuli with different features. As a interpretive point, however, it is important to emphasize that these differing activations can only be understood by the relative contribution that they make to a single network for self-projection.

The link between moral deliberation and self-projection may not come as a surprise. Moral dilemmas may engage the core self-projection network in a trivial sense, as considering them requires imagining performing actions, and so self-projection. However, not all studies of moral decision-making require imaginative self-projection: some require judging the blameworthiness of others, others just the detection of morally salient features of a situation. I suggest that the better analogy might be with social deliberation.

Here is an (admittedly speculative) proposal. Moral decisions are bound up with thinking about ourselves as part of a moral community. To judge our own actions as good or bad is (in part) to judge whether others have reason to praise or blame us. To judge others is (in part) to determine their attitudes towards other members of the moral community—contempt for oth-

Study	Contrast	PC/PCC	TPJ	mFC/OFC	MTL
[57]	viewing moral> unpleasant scenarios		*	*	*
[57]	viewing moral> neutral scenarios	*	*	*	*
[55]	moral> semantic judgments	*	*	*	*
[29]	moral> nonmoral		*	*	
[61]	moral> sociomoral or neutral norm violations			*	*
[60]	detection of moral> neutral information	*	*	*	
[60]	detection of moral> strategic information	*	*	*	
[37]	sociomoral> pathogen/nonmoral scenarios	*	*	*	*
[41]	main effect of moral vs nonmoral deliberation	*	*	*	†
[56]	sociomoral > grammatical judgments		*	*	*
[42]	main effect of moral vs nonmoral deliberation	*	*	*	†
[95]	(implicit or explicit) moral > nonmoral	*	*	*	*

Table 1 Networks involved in moral vs nonmoral decision-making. A > indicates greater activation in the former condition relative to the latter. A † indicates that the area was not among planned ROIs for the experiment, and so no information about activation was presented.

ers deserves blame, while sincere intentions might not. To judge our own action as wrong is in part to judge ourselves as having made a breach in our relationships with others in the moral community, and seek to repair it.[21] And (as Derek Baker has helpfully emphasized to me), to make *any* moral judgment is also, in part, to judge that everyone else has reason to make the same judgment, for moral judgments are universalizable in a way that many merely social judgments are not.

All of these judgments involve not just social cognition, but especially tricky forms of *higher-order* social cognition: to determine whether an action would make us blameworthy or merely impertinent, we might have to determine what others would justifiably think about what we were thinking when we decided to steal the policeman's cap. Though higher-order, this sort of deliberation is still recognizable as an iterated form self-projection: we must imagine the world from a variety of viewpoints, and relate those viewpoints back to our own.

That moral decision-making is a special kind of social cognition is an old idea.[22] The present proposal is sketched in broad enough strokes to be compatible with a variety of ethical and metaethical positions, though further empirical research (especially on the mPFC) might support a more precisely drawn story. For present purposes, however, note that this remains a *single-track*, not a dual-track, theory. Moral decision-making is the result of a single process: higher-order social deliberation. This single process may result in contradictory issuances, which then need to be sorted out (we might decide that, from one perspective we are blameworthy, from another charming, and then have to adjudicate which is best). This process may even have blind spots that systematically mislead us in certain cases.[23] But that would be a single process, instantiated by a single, complex brain network.

5 Recap and Conclusion

I conclude by way of a recap, to situate the argument of this paper within the larger context with which I started. I began with the observation that people appear to have divergent intuitions in the face of ethical dilemmas like the Trolley and Footbridge problem, and that these divergent intuitions can lead to a sense of internal conflict. Dual-track theories present themselves as the best explanation for these divergent intuitions. fMRI evidence was supposed to provide evidence for the dual track theory, insofar as the best explanation of the neuroimaging data involved two distinct mechanisms responsible for different sorts of moral intuitions.

I gave reason to think that an influential version of the dual-track interpretation of the neuroimaging data, on which dilemmas activated distinct emotional or cognitive pathways, wasn't well-supported by the neuroimaging data: in each case, activated areas were specific neither to cognition nor emotion. I then gave reason to think that the mere presence of distinct patterns of activation did not favor any dual-track theory.

[21] For a development of this idea, see chapter 4 of [97]. In an intriguing study, Finger et al. noted a common substrate in dmPFC for moral transgressions and witnessed social transgressions [61]. They suggest that this may correspond to the intention to repair social relationships with others, a usual requirement of such transgressions. This further step in moral decision-making deserves careful study.

[22] A canonical source is [1]. For a contemporary proposal that also draws on neuroimaging evidence, see Section 2 of [7].

[23] For example, insofar as we have moral intuitions about particular cases, they may be shaped by how easy or difficult it is to form higher-order judgments. Adam Smith, for example, notes that "if we consider all the different passions of human nature, we shall find that they are regarded as decent, or indecent, just in proportion as mankind are more or less disposed to sympathize with them" ([1] I.ii).

For one, both single- and dual-track theories should result in patterns of activation that differed across dilemma-types. For another, I argued that an area-by-area interpretation of brain activation was unlikely to provide compelling evidence for any cognitive theory, as the many-many mappings between brain areas and cognitive functions made for a combinatorial explosion of possible explanations of observed activity. I argued instead for an alternative, network-based, strategy of interpretation, and then gave a network-based analysis that supported a single-track model. This single-track model, on which moral decision-making was a species of self-projection, was both well-supported by the data and fit with plausible metaethical positions about the nature and goal of moral decision-making. Thus, we have reason to think that neuroimaging data supports a single-track theory over a dual-track one.

This is not a knockdown argument for a single-track view of moral decision-making. First, as I noted from the outset, neuroimaging data is only one source of evidence for a theory. I have not touched upon other evidence that might be relevant. Second, neuroimaging data does not trump all. It is tempting to suppose that fMRI provides a shortcut around the messy methodological questions that plague other sorts of model-building in cognitive science. If I have shown nothing else, I hope that I have shown this to be false: the interpretation of neuroimaging data is as methodologically complex as the interpretation of any other psychological data. Third, it is open question whether a cross-domain abductive analysis that divides the neuroimaging data into two distinct functional tracks might not be done, and look more compelling than my single-track interpretation. (I doubt that such an analysis can be given, but failures of imagination are rarely the end of the story). Fourth, I did not consider the possibility that moral decision-making might be an example of a partially *degenerate* cognitive function: that is, a function that is implemented by partially overlapping but distinct and largely redundant brain networks. Degeneracy is a common feature of many biological systems [98], and is increasingly thought to be an important feature of a variety of cognitive systems [99]. That said, a degenerate model will have more in common *philosophically* with a single-track than a dual-track model, for it does not predict the presence of widespread error within any particular track. Fifth and finally, the proposal sketched is a speculative taxonomic hypothesis: its primary purpose was to argue that there is decent evidence for one track rather than two. Fleshing out the functional story by teasing out the differential contributions of particular regions within the network would require substantial further work, and linking them to the philosophical story I proposed would take further work. Nevertheless, as I have indicated above, there are both empirical and philosophical single-track accounts of moral reasoning, and the model proposed might well be assimilated to work already done by those accounts.

These caveats should not diminish the force of my argument, however. Science is a matter of collecting evidence for and against hypotheses. When the tally is completed, neuroimaging data should belong in the column against dual-track theory. The connection between self-projection and moral decision-making suggests further tests, and requires refinement in numerous ways. That is a good thing: we want our hypotheses to suggest fruitful research projects, and this version of the single-track theory fits the bill.

When thinking about future research, it is worth taking a further step back. Both single- and dual-track theories were interesting because they joined empirical content to philosophical theses. The philosophical commitments of both theories are also worth keeping in mind even while doing empirical work. The dual-track theory was motivated by an apparent inconsistency between differing responses to moral dilemmas. Moral errors were thus explained by the presence of mechanisms that are indifferent to the rightness or wrongness of action. Single-track theories, in contrast, explain error by appeal to a single process that is correct in some contexts, misguided in others, but not inherently flawed. This suggests that the best taxonomy of mental mechanisms (an empirical claim) might depend in part on what we think the right moral theory should be (a philosophical claim). Only given a good theory of moral success and failure can we interpret mental mechanisms as inherently or contingently flawed. Thus, philosophers may have as much to contribute to the analysis of experimental results as experimental results have to contribute to philosophy.

Thanks to Jennifer Ashton, Derek Baker, Selim Berker, Tom Dougherty, Dave Hilbert, Esther Klein, Ruth Leys, Tristram McPherson, Chris Mole, Sally Sedgwick, Nick Stang, Rachel Zuckert, and the 2009-10 UIC Humanities Institute fellows for helpful comments and discussions. I received useful feedback on previous drafts from audiences at University of Miami, Johns Hopkins University, and University of Illinois at Chicago. The present work was supported in part by a fellowship from the UIC Institute for the Humanities.

References

1. Adam Smith. *The Theory of Moral Sentiments.* Liberty Fund, Indianapolis, 1982.

2. Joshua D Greene, Leigh E Nystrom, Andrew D Engell, John M Darley, and Jonathan D Cohen. The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400, Oct 2004.
3. Selim Berker. The normative insignificance of neuroscience. *Philosophy & Public Affairs*, 37(4):293–329, 2009.
4. J D Greene, R B Sommerville, L E Nystrom, J M Darley, and J D Cohen. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108, Sep 2001.
5. Joshua D Greene. The secret joke of Kant's soul. In Sinnott-Armstrong [100], pages 35–80.
6. P. Singer. Ethics and intuitions. *The Journal of Ethics*, 9(3):331–352, 2005.
7. J. Allman and J. Woodward. What are moral intuitions and why should we care about them: a neurobiological perspective. *Philosophical Issues*, 18(1):164–185, 2008.
8. FM Kamm. Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs*, 37(4):330–345, 2009.
9. Richard Henson. What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology*, 58(2):193–233, Feb 2005.
10. Russell A Poldrack. The role of fMRI in cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology*, 18(2):223–227, Apr 2008.
11. C.J. Price and K.J. Friston. Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, 22(3):262–275, 2005.
12. Russell A Poldrack. Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2):59–63, Feb 2006.
13. R.J. Maddock. The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain. *Trends in Neurosciences*, 22(7):310–316, 1999.
14. Hedy Kober, Lisa Feldman Barrett, Josh Joseph, Eliza Bliss-Moreau, Kristen Lindquist, and Tor D Wager. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*, 42(2):998–1031, Aug 2008.
15. A.E. Cavanna and M.R. Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564, 2006.
16. B.A. Vogt and S. Laureys. Posterior cingulate, precuneal & retrosplenial cortices: Cytology & components of the neural network correlates of consciousness. *Progress in brain research*, 150:205, 2005.
17. K. Vogeley and G.R. Fink. Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences*, 7(1):38–42, 2003.
18. K. Vogeley, M. May, A. Ritzl, P. Falkai, K. Zilles, and GR Fink. Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, 16(5):817–827, 2004.
19. P.L. Jackson, E. Brunet, A.N. Meltzoff, and J. Decety. Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia*, 44(5):752–761, 2006.
20. R. Saxe and N. Kanwisher. People thinking about thinking people the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4):1835–1842, 2003.
21. K.N. Ochsner, K. Knierim, D.H. Ludlow, J. Hanelin, T. Ramachandran, G. Glover, and S.C. Mackey. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, 16(10):1746–1772, 2004.
22. J. Okuda, T. Fujii, H. Ohtake, T. Tsukiura, K. Tanji, K. Suzuki, R. Kawashima, H. Fukuda, M. Itoh, and A. Yamadori. Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *Neuroimage*, 19(4):1369–1380, 2003.
23. D.R. Addis, A.T. Wong, and D.L. Schacter. Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45(7):1363–1377, 2007.
24. M.K. Johnson, C.L. Raye, K.J. Mitchell, S.R. Touryan, E.J. Greene, and S. Nolen-Hoeksema. Dissociating medial frontal and posterior cingulate activity during self-reflection. *Social Cognitive and Affective Neuroscience*, 1(1):56, 2006.
25. K.N. Ochsner, J.S. Beer, E.R. Robertson, J.C. Cooper, J.D.E. Gabrieli, J.F. Kihsltrom, and M. D'Esposito. The neural correlates of direct and reflected self-knowledge. *Neuroimage*, 28(4):797–814, 2005.
26. P. Fransson and G. Marrelec. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: evidence from a partial correlation network analysis. *Neuroimage*, 42(3):1178–1184, 2008.
27. D.A. Gusnard, E. Akbudak, G.L. Shulman, and M.E. Raichle. Medial prefrontal cortex and

self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(7):4259–4264, 2001.
28. Seralynne D Vann, John P Aggleton, and Eleanor A Maguire. What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, 10(792-802), 2009.
29. J.S. Borg, C. Hynes, J. Van Horn, S. Grafton, and W. Sinnott-Armstrong. Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, 18(5):803–817, 2006.
30. G. Hein and R.T. Knight. Superior temporal sulcus—It's my area: Or is it? *Journal of Cognitive Neuroscience*, 20(12):2125–2136, 2008.
31. A.D. Engell and J.V. Haxby. Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia*, 45(14):3234–3241, 2007.
32. S. Materna, P.W. Dicke, and P. Thier. The posterior superior temporal sulcus is involved in social communication not specific for the eyes. *Neuropsychologia*, 46(11):2759–2765, 2008.
33. E. Redcay. The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. *Neuroscience and Biobehavioral Reviews*, 32(1):123–142, 2008.
34. S.M. Kosslyn, L.M. Shin, W.L. Thompson, R.J. McNally, S.L. Rauch, R.K. Pitman, and N.M. Alpert. Neural effects of visualizing and perceiving aversive stimuli: a PET investigation. *Neuroreport*, 7(10):1569, 1996.
35. E.M. Reiman, R.D. Lane, G.L. Ahern, G.E. Schwartz, R.J. Davidson, K.J. Friston, L.S. Yun, and K. Chen. Neuroanatomical correlates of externally and internally generated human emotion. *American Journal of Psychiatry*, 154(7):918–925, 1997.
36. T.D. Wager and E.E. Smith. Neuroimaging studies of working memory: a meta-analysis. *Cogn Affect Behav Neurosci*, 3(4):255–274, 2003.
37. J.S. Borg, D. Lieberman, and K.A. Kiehl. Infection, incest, and iniquity: Investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, 20(9):1529–1546, 2008.
38. J. Haidt. The emotional dog and its rational tail. *Psychological Review*, 108(4):814–834, 2001.
39. R. Saxe and A. Wexler. Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43(10):1391–1399, 2005.
40. D. Samson, I.A. Apperly, C. Chiavarino, and G.W. Humphreys. Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience*, 7(5):499–500, 2004.
41. L. Young and R. Saxe. The neural basis of belief encoding and integration in moral judgment. *Neuroimage*, 40(4):1912–1920, 2008.
42. L. Young and R. Saxe. An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, 21(7):1396–1405, 2009.
43. A. Bechara and A.R. Damasio. The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, 52(2):336–372, 2005.
44. J.D. Greene. Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8):322–323, 2007.
45. Guy Kahane and Nicholas Shackel. Do abnormal responses show utilitarian bias? *Nature*, 452(7185):E5, Mar 2008.
46. Michael Koenigs and Daniel Tranel. Irrational economic decision-making after ventromedial prefrontal damage: evidence from the ultimatum game. *Journal of Neuroscience*, 27(4):951–956, Jan 2007.
47. Michael Koenigs, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138):908–911, Apr 2007.
48. J. Moll and R. de Oliveira-Souza. Response to Greene: Moral sentiments and reason: friends or foes? *Trends in Cognitive Sciences*, 11(8):323–324, 2007.
49. A.R. Damasio. *Descartes' Error: Emotion, reason, and the human brain*. GP Putnam's Sons, New York, 1994.
50. David Badre and Mark D'Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–69, Sep 2009.
51. Christopher Kennard Parashkev Nachev and Masud Husain. The functional anatomy of the frontal lobes. *Nature Reviews Neuroscience*, 10(829), 2009.
52. D.M. Amodio and C.D. Frith. Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4):268–277, 2006.
53. Todd A Hare, Colin F Camerer, and Antonio Rangel. Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, 324(5927):646–648, May 2009.
54. J.W. Kable and P.W. Glimcher. The neural correlates of subjective value during intertemporal

choice. *Nature Neuroscience*, 10(12):1625–1633, 2007.

55. H.R. Heekeren, I. Wartenburger, H. Schmidt, K. Prehn, H.P. Schwintowski, and A. Villringer. Influence of bodily harm on neural correlates of semantic and moral decision-making. *Neuroimage*, 24(3):887–897, 2005.
56. K. Prehn, I. Wartenburger, K. Meriau, C. Scheibe, O.R. Goodenough, A. Villringer, E. van der Meer, and H.R. Heekeren. Individual differences in moral judgment competence influence neural correlates of socio-normative judgments. *Social Cognitive and Affective Neuroscience*, 3(1):33, 2008.
57. J. Moll, R. de Oliveira-Souza, P.J. Eslinger, I.E. Bramati, J. Mourao-Miranda, P.A. Andreiuolo, and L. Pessoa. The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22(7):2730, 2002.
58. J. Moll, R. de Oliveira-Souza, G.J. Garrido, I.E. Bramati, E.M.A. Caparelli-Daquer, M.L.M.F. Paiva, R. Zahn, and J. Grafman. The self as a moral agent: linking the neural bases of social agency and moral sensitivity. *Social Neuroscience*, 2(3):336–352, 2007.
59. J.P. Mitchell, M.R. Banaji, and C.N. MacRae. The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 17(8):1306–1315, 2005.
60. D. Robertson, J. Snarey, O. Ousley, K. Harenski, F.D.B. Bowman, R. Gilkey, and C. Kilts. The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia*, 45(4):755–766, 2007.
61. E.C. Finger, A.A. Marsh, N. Kamel, D.G.V. Mitchell, and J.R. Blair. Caught in the act: The impact of audience on the neural response to morally and socially inappropriate behavior. *NeuroImage*, 33(1):414–421, 2006.
62. Kent A Kiehl. Without morals: The cognitive neuroscience of criminal psychopaths. In Sinnott-Armstrong [100], pages 119–150.
63. Ricardo de Oliveira-Souza, Fátima Azavedo Ignácio, and Jorge Moll. The antisocials among us. In Sinnott-Armstrong [100], pages 151–158.
64. R.J. Compton, M.T. Banich, A. Mohanty, M.P. Milham, J. Herrington, G.A. Miller, P.E. Scalf, A. Webb, and W. Heller. Paying attention to emotion: An fmri investigation of cognitive and emotional stroop tasks. *Cogn Affect Behav Neurosci*, 3(2):81–96, 2003.
65. A. Etkin, T. Egner, D.M. Peraza, E.R. Kandel, and J. Hirsch. Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, 51(6):871–882, 2006.
66. C.A. Boettiger and M. D'Esposito. Frontal networks for learning and executing arbitrary stimulus-response associations. *Journal of Neuroscience*, 25(10):2723, 2005.
67. Silvia A. Bunge. How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, 4(4):564–579, 2004.
68. Luiz Pessoa. On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–58, Feb 2008.
69. J.D. Greene, S.A. Morelli, K. Lowenberg, L.E. Nystrom, and J.D. Cohen. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3):1144–1154, 2008.
70. Chris Mole and Colin Klein. Confirmation, refutation and the evidence of fMRI. In Stephen José Hanson and Martin Bunzl, editors, *Foundational Issues of Human Brain Mapping*, page : (forthcoming). MIT Press, Cambridge, 2010.
71. AD Craig. How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, 3(655-666), 2002.
72. AD Craig. Interoception and emotion: a neuroanatomical perspective. In Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of emotion*, pages 272–288. The Guilford Press, New York, 2008.
73. Alan G Sanfey, James K Rilling, Jessica A Aronson, Leigh E Nystrom, and Jonathan D Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758, Jun 2003.
74. D.A. Kravitz and S. Gunto. Decisions and perceptions of recipients in ultimatum bargaining games. *Journal of Socio-Economics*, 21(1):65–84, 1992.
75. Toshio Yamagishi, Yutaka Horita, Haruto Takagishi, Mizuho Shinada, Shigehito Tanida, and Karen S Cook. The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences*, 106(28):11520–11523, 2009.
76. Nicola Grahek. *Feeling pain and being in pain.* The MIT Press, Cambridge, MA, 2007.
77. L.F. Barrett and E. Bliss-Moreau. Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41:167–218, 2009.

78. S.R. Quartz. Reason, emotion and decision-making: risk and reward computation with feeling. *Trends in Cognitive Sciences*, 13(5):209–215, 2009.
79. S. Duncan and L.F. Barrett. Affect is a form of cognition: A neurobiological analysis. *Cognition & emotion*, 21(6):1184, 2007.
80. C. Camerer, G. Loewenstein, and D. Prelec. Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, 43(1):9–64, 2005.
81. C. Mole, C. Kubatzky, J. Plate, R. Waller, M. Dobbs, and M. Nardone. Faces and Brains: The Limitations of Brain Scanning in Cognitive Science. *Philosophical Psychology*, 20(2):197–207, 2007.
82. Jorge Moll and Ricardo de Oliveira-Souza. Moral judgments, emotions and the utilitarian brain. *Trends Cogn Sci*, 11(8):319–321, Aug 2007.
83. J. Moll, R. Zahn, R. de Oliveira-Souza, F. Krueger, and J. Grafman. The neural basis of human moral cognition. *Nature Reviews Neuroscience*, 6(10):799–809, 2005.
84. Max Coltheart. What has functional neuroimaging told us about the mind (so far)? *Cortex*, 42(3):323–331, Apr 2006.
85. P. Indefrey. A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language Learning*, 56:279, 2006.
86. Paul E Meehl. Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, 46:806–834, 1978.
87. Colin Klein. Images are not the evidence of neuroimaging. *British Journal for the Philosophy of Science*, page : (forthcoming), 2010.
88. Robert L Savoy. History and future directions of human brain mapping and functional imaging. *Acta Psychologica*, 107:9–42, 2001.
89. M.L. Anderson. The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology*, 20(2):143–174, 2007.
90. R.L. Buckner and D.C. Carroll. Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57, 2007.
91. D.L. Schacter, D.R. Addis, and R.L. Buckner. Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661, 2007.
92. DH Ingvar. "Memory of the future": An essay on the temporal organization of conscious awareness. *Human neurobiology*, 4(3):127, 1985.
93. E. Tulving. Memory and consciousness. *Canadian Psychology*, 26(1):1–12, 1985.
94. B. Levine, M. Freedman, D. Dawson, S. Black, and D.T. Stuss. Ventral frontal contribution to self-regulation: Convergence of episodic memory and inhibition. *Neurocase*, 5(3):263–275, 1999.
95. C.L. Harenski, O. Antonenko, M.S. Shane, and K.A. Kiehl. A functional imaging investigation of moral deliberation and moral intuition. *Neuroimage*, 49:2707–2716, 2009.
96. L. Young, J.A. Camprodon, M. Hauser, A. Pascual-Leone, and R. Saxe. 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, 2010.
97. TM Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press, Cambridge, 2008.
98. G.M. Edelman and J.A. Gally. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24):13763, 2001.
99. K.J. Friston and C.J. Price. Degeneracy and redundancy in cognitive anatomy. *Trends in Cognitive Sciences*, 7(4):151–152, 2003.
100. Walter Sinnott-Armstrong, editor. *Moral Psychology*, volume 3. MIT Press, Cambridge, 2008.