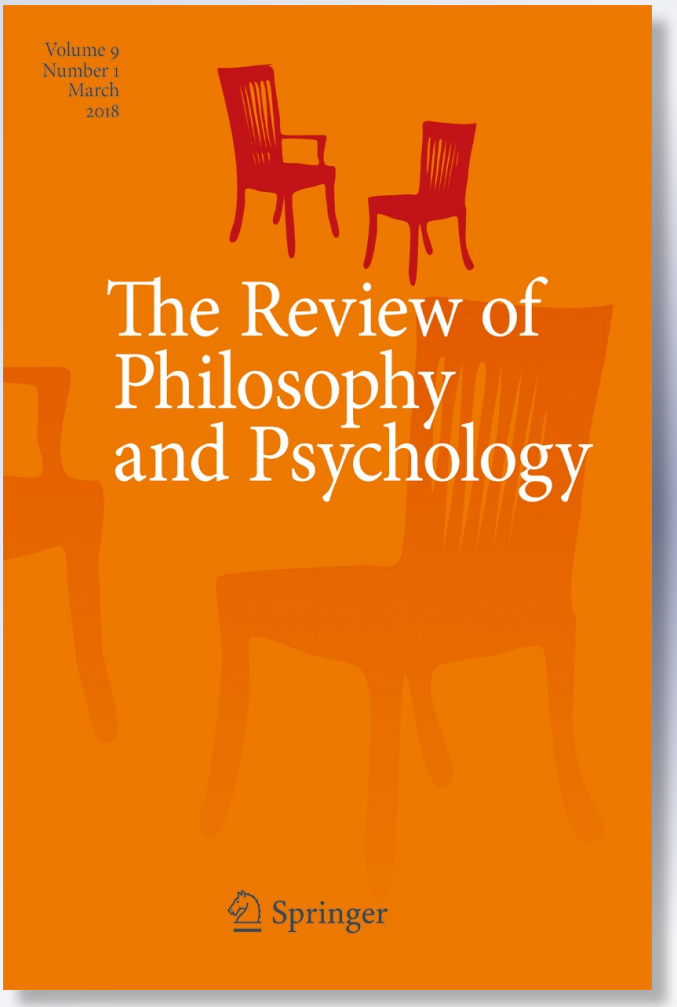
Mapping Cognitive Structure onto the Landscape of Philosophical Debate: an Empirical Framework with Relevance to Problems of Consciousness, Free will and Ethics

Jared P. Friedman & Anthony I. Jack

Review of Philosophy and Psychology

ISSN 1878-5158
Volume 9
Number 1

Rev.Phil.Psych. (2018) 9:73-113
DOI 10.1007/s13164-017-0351-6



Springer

Rev.Phil.Psych. (2018) 9:73–113
DOI 10.1007/s13164-017-0351-6

Mapping Cognitive Structure onto the Landscape of Philosophical Debate: an Empirical Framework with Relevance to Problems of Consciousness, Free will and Ethics

Jared P. Friedman [1,2,3] · **Anthony I. Jack** [1,2,3,4,5,6]

Published online: 1 July 2017

Abstract There has been considerable debate in the literature as to whether work in experimental philosophy (X-Phi) actually makes any significant contribution to philosophy. One stated view is that many X-Phi projects, notwithstanding their focus on topics relevant to philosophy, contribute little to philosophical thought. Instead, it has been claimed the contribution they make appears to be to cognitive science. In contrast to this view, here we argue that at least one approach to X-Phi makes a contribution which parallels, and also extends, historically salient forms of philosophical analysis, especially contributions from Immanuel Kant, William James, Peter F. Strawson and Thomas Nagel. The framework elaborated here synthesizes philosophical theory with empirical evidence from psychology and neuroscience and applies it to three perennial philosophical problems. According to this account, the origin of these three problems can be illuminated by viewing them as arising from a tension between two distinct types of cognition, each of which is associated with anatomically independent and functionally inhibitory neural networks. If the parallel we draw, between an empirical project and historically highly influential examples of philosophical analysis, is viewed as convincing, it follows that work in the cognitive sciences can contribute directly to

✉ Jared P. Friedman
Jpf66@case.edu

[1] Department of Philosophy, Case Western Reserve University, Cleveland, OH, USA

[2] Inamori International Center for Ethics and Excellence, Case Western Reserve University, Cleveland, OH, USA

[3] Department of Organizational Behavior, Case Western Reserve University, Cleveland, OH, USA

[4] Department of Neurology, Case Western Reserve University, Cleveland, OH, USA

[5] Department of Neuroscience, Case Western Reserve University, Cleveland, OH, USA

[6] Department of Psychology, Case Western Reserve University, Cleveland, OH, USA

philosophy. Further, this conclusion holds whether the empirical details of the account are correct or not.

> "*This method of watching or even occasioning a contest between [mutually exclusive metaphysical] assertions, not in order to decide it to the advantage of one party or the other, but to investigate whether the object of the dispute is not perhaps a mere mirage at which each would snatch in vain without being able to gain anything even if he met with no resistance – this procedure, I say, can be called the* ***skeptical method****...[T]he skeptical method aims…to discover the point of misunderstanding in disputes that are honestly intended and conducted with intelligence by both sides…*" (Kant 1787/1998:A423/B451-A424/B452; emphasis in original)

> "*That the human mind will ever give up metaphysical researches is as little to be expected as that we, to avoid inhaling impure air, should prefer to give up breathing all together. There will, therefore, always be metaphysics in the world; nay, everyone, especially every man of reflection, will have it and, for want of a recognized standard, will shape it for himself after his own pattern.*" (Kant 1902, p. 367)

> "*'Thoughts' and 'things' are names for two sorts of object, which common sense will always find contrasted and will always practically oppose to each other. Philosophy, reflecting on the contrast, has varied in the past in her explanations of it, and may be expected to vary in the future.*" (James 1904, p.477)

1 Introduction

There are some seemingly intractable questions that have remained at the heart of philosophical discourse since they were first asked. Is the mind distinct from the brain, or are we just physical stuff? Are we autonomous agents, or merely at the mercy of the causal and mechanistic laws of nature? When, if ever, is it acceptable to sacrifice one life for the greater good of many? That these questions have remained at the heart of philosophy for so long, and that their proposed 'solutions' (e.g., monism vs. dualism, compatibilism vs. incompatibilism, utilitarianism vs. deontology) appear linked to incommensurable perspectives, strikes us as enigmatic. Might the intractable nature of these issues reflect something peculiar about *us* rather than something peculiar about the way the world *is*? Put another way, do these intractable debates originate from how we think about these issues, rather than reflecting a feature inherent to the phenomena we seek to apprehend (e.g., mind/brain, human agency, and normative truth)? If so, it might be that competition between distinct cognitive processes is what renders these three philosophical issues problematic. Suppose not only that this is so, but also that these cognitive processes arise from a basic feature of our evolved neurobiology. If so, it follows that our neural architecture might not only inevitably create these schisms in

perspective, but also prevent an ultimate solution that decides in favor of one competing view over another. In particular, it may never be possible to decide between the competing views that, each in different ways, attempt to resolve the above three problems of interest. We think this is this case, and will argue that this conclusion is supported not only by some insightful and highly influential examples of philosophical analysis, but also independently by empirical evidence from the cognitive sciences.

In this essay, we focus on an account of cognitive structure which is informed by, and makes predictions relating to, three overlapping research traditions in the cognitive sciences: Neuropsychology, Cognitive Neuroscience and Psychology (discussed in more detail in Section 2.1. See also Fig. 1, left column[1]). Central to this account is the discovery (Shulman et al. 1997b), and more recent cognitive characterization (Jack et al. 2012), of an antagonistic relationship between two large-scale cortical networks. These networks are known, for historical reasons,[2] as the task positive network (TPN) and the default mode network (DMN) (Fox et al. 2005). Work in cognitive neuroscience suggests that the TPN is associated with various types of non-social cognition, which can be broadly characterized as 'analytic' in nature – including logical, mathematical, empirical and critical reasoning. In contrast, the DMN is associated with various types of social-emotional cognition, which can be broadly characterized as 'empathic' in nature – including thinking about one's own and other's mental states (see section 2.1 for a review). Critically, these networks have a distinctive relationship insofar as they generally tend to suppress each other (although, as discussed below, they can also work together for some tasks). Activity in these networks has been likened to a see-saw, because when one network is 'up' (activated above resting baseline) the other network is usually 'down' (deactivated below resting baseline). We claim that the antagonistic relationship between these two brain networks maps onto, and hence can partially explain, the philosophical schisms discussed here.

The mapping between this neural divide and the aforementioned philosophical schisms can be best understood by reference to an intermediate level of description (discussed in more detail in Section 2.2. See also Fig. 1, right column). Following Dennett's (1989) framework,[3] we identify three 'cognitive stances' which are of particular relevance to work in philosophy (Robbins and Jack 2006). These are the Physical stance, the Intentional stance and the Phenomenal stance. Each of these stances is understood to be realized by different configurations of the two brain networks. At one extreme of the see-saw, the Physical stance corresponds to engagement of the TPN and suppression of the DMN. At the other extreme, the Phenomenal stance corresponds to engagement of the DMN and suppression of the TPN. The

[1] By Neuropsychology, we mean the study of cognition as informed by studying individuals with neurological deficits (either developmental or acquired). The work in Cognitive Neuroscience that informs this account derives primarily from brain imaging and from network modelling. The work in Psychology that primarily informs the account derives from measures that assess individual differences in social and non-social cognitive processing.

[2] It was initially thought the key cognitive distinction between the TPN and DMN was whether individuals were engaged in a task or 'defaulting' to undirected spontaneous cognition. As will become apparent, this cognitive characterization is incorrect. However, the names used to refer to the networks are now well established.

[3] Dennett might be said to have identified a level of description which mediates between psychological and philosophical characterizations of distinctive ways of thinking. Unsurprisingly, this level of description suits current purposes well.

Fig. 1 A schematic depiction of mental structure and relationships to philosophical issues. Left column illustrates different types of cognition and relationship to Neuropsychological conditions, brain imaging, and individual difference variables. Right column relates cognitive structure to philosophically relevant ways of thinking and conceptual tensions. For all figures, the red barbells between the blue and pink circles represent mutual antagonism between cognitive processes and abstracted representations associated with each. The smaller red barbells and double sided green arrows connecting the green and blue circles, and the green and pink circles, represent limited forms of antagonism and compatibility between the cognitive processes and abstracted representations associated with each. **a** (top left): An updated and broader characterization of the cognitive processes associated with the Physical, Phenomenal and Intentional stances. The rectangular boxes indicate individual difference measures thought to assess the individual's tendency and/or ability to engage each type of thinking. **b** (middle left): Relationships to deficits associated with Neuropsychological disorders, as described by Robbins and Jack (2006). **c** (lower left): Mapping the Physical, Phenomenal and Intentional stances onto different configurations of recruitment of the Default Mode Network (DMN) and Task Positive Network (TPN), adapted from Jack et al. (2013). **d** (top right): Robbins and Jack's (2006) mapping of the stances onto Brentano's problem and the problem of consciousness. Brentano's problem emerges from the tension between the Physical and Intentional stances whereas the more salient problem of consciousness emerges from the greater tension between the Physical and Phenomenal stances. **e**: A novel proposed mapping the stances onto the problem of determinism as articulated by Strawson. **f**: Mapping the stances onto the debate between utilitarian and deontological ethics, as proposed by Jack et al. 2014. The loops indicate our hypothesis that Utilitarian ethics (Mill) reflects a blend between cognitive processes associated with the Physical and Intentional stances, whereas Deontological ethics (Kant) reflects a blend between the Phenomenal and Intentional stances. The cognitive processes and attitudes in each circle are also proposed to reflect Kant's notions of theoretical and practical reason

Intentional stance corresponds to a coalition that draws on neural resources in both the TPN and DMN, and demonstrates both overlap and tension with each of the other stances. Note that while these three stances capture the modes of engagement of these networks that are most pertinent to the philosophical questions under consideration, we are not claiming they capture the full breadth of cognitive processing associated with each network configuration. For instance, adoption of the Physical Stance engages the TPN and suppresses the DMN; and so do other analytic tasks, such as logical and mathematical reasoning, focused attention and non-social working memory tasks.

We trace each of the three philosophical schisms discussed here to tensions between these stances, or in the case of ethics, to two viewpoints which combine these stances in different ways (see Fig. 1, right column). The empirical details of this account, and initial descriptions of the proposed mappings, are covered in greater detail in the next section. However, before we delve more deeply into these details, it will be helpful to better clarify the overall objective of this essay and its structure.

This essay's primary thesis is that the empirical work reviewed here informs and enlightens the three philosophical issues under discussion in a manner which is no less philosophically relevant and substantive than the historically influential examples of conceptual analysis we discuss. Hence, a key conclusion of this essay is that work in cognitive science can, under the right conditions, parallel certain types of philosophical analysis so closely that it reveals an area of overlap between the disciplines. Therefore, our claim is not that the current project is interdisciplinary in the weak sense that the work relates to and/or makes use of findings, concepts or methods from more than one discipline. Instead, our claim is that the current project is interdisciplinary in the stronger sense that it both belongs to and constitutes an example of work in each of the disciplines.

This is, of course, not the first time such an idea has come to light. Since its inception, Experimental Philosophy (X-Phi) offered the apparent promise of establishing an area of overlap between Philosophy and the Cognitive Sciences. However, as interest in X-Phi has grown, this initial promise has also come under closer scrutiny (e.g., Kauppinen 2007). As a result, not only critics of X-Phi, but also some of its best-known practitioners and strongest proponents (e.g. Joshua Knobe and Edouard Machery), have come to question the degree to which most existing work in X-Phi is directly relevant to philosophy (see section 4). These analyses are insightful and important, however we argue that any general conclusion about the Philosophical import of work in X-Phi would be premature, since the project we describe here differs from (the majority of) other X-Phi projects in a number of important respects. We will discuss these differences in more detail in Section 4. For now, we focus on the first, and in many ways the most important difference, namely that our approach differs in terms of how it approaches the philosophy.

Much work in philosophy aims at one or other position (e.g., physicalism or dualism), and then uses conceptual analysis to tackle objections and/or find support for the position. Similarly, most X-Phi projects are aimed at supporting one or another position, usually by examining the intuitions that drive this or the contrary position. A recent critique by Knobe (2016) is aimed at these types of projects, and questions whether this approach of examining intuitions can directly inform conceptual analysis. We view Knobe's (2016) argument as cogent, but orthogonal to the project we advance here.

Unlike most work in X-Phi, we find it more appealing to approach metaphysical debates by assuming that each position has its own distinct, and respectable, motivations. This is indeed suggested by the presence of such disparate perspectives both among professional philosophers and the folk. Hence, our approach is neither aimed at supporting a particular position, nor does it examine intuitions in order to find support for, or undermine, the supposed underlying basis of arguments based on conceptual analysis. Conceptual analysis, as usually understood, is not our target. Rather, the empirical approach we discuss here was both inspired by, and aims to inform, a different type of philosophical analysis. We dub this type of philosophical analysis the 'illumination of conceptual tension' (Section 3). Similar to Kant's skeptical method (see introductory quote, and Section 3.3), the idea is not to find evidence that advantages one metaphysical point of view over another, but to try to uncover the origins of the tension between different metaphysical viewpoints.

Wearing our 'cognitive scientist' hats, we have interpreted the existence of long-running and unresolved disputes in philosophy as anecdotal evidence suggesting the existence of a tension between different psychological faculties for understanding. We operationalized this as a hypothesis, called the 'opposing domains hypothesis', which we then went out to test experimentally using the established methods of cognitive science (Jack et al. 2012, 2014, 2016; Jack and Robbins 2012; Jack 2014). It is important to note that while the testing of such a hypothesis may be most accurately described as work in cognitive science, the generation of such a hypothesis is a recognizable move in philosophy. Indeed, this initial move was anticipated by the influential philosophers whose work we review here – they simply did not have the methods (or, perhaps, the inclination) to go on to test their hypotheses empirically. As we detail throughout Section 3, each of them suggests, at some point in their writings, that philosophical tensions arise from our psychology: Kant talks about antinomies that arise from the relationship between different cognitive faculties that we impose upon our experience in order to make sense of it. James claims that the history (and future) of philosophy reflects a clash of psychological temperaments (i.e., personality differences). Strawson refers to a tension that occurs 'in us' between participant and objective attitudes. And Nagel suggests that the appearance of a contingent relationship between experiential and physical states may occur because we possess 'disparate types of imagination'.

Adopting our 'philosopher' hats, our goal in this essay is to show the reader how empirical investigations into hypotheses of this nature both closely parallel and extend the philosophical analyses offered by these influential thinkers. To accomplish this aim, the remainder of this essay is structured as follows. In section 2, we outline a basic mapping from cognitive structure to the three philosophical schisms of interest. Specifically, in section 2.1, we provide a brief review of the empirical evidence which informs the relevant claims about cognitive structure, and in section 2.2 we outline how our cognitive structure maps onto three philosophically relevant stances, and from there onto the three philosophical schisms of interest. In section 3, we explicitly draw similarities between the philosophical import afforded by this empirical project and the more traditional, and highly influential, types of philosophical analysis it parallels. Specifically, in separate sections we tackle influential contributions by Thomas Nagel, Peter F. Strawson, and Immanuel Kant. We then go on to discuss recent work on the conflict between utilitarian and deontological ethics. In section 4, we focus on the

traditional domain of X-Phi, namely relating psychological factors to differences in philosophical intuition and belief. In this section we differentiate our project from other projects in experimental philosophy, and note a close parallel between some of our X-Phi findings and remarks by William James about the underlying cause of disagreements in philosophy. Section 5 is the summary and conclusion.

2 From Cognitive Structure to Schisms in Understanding

2.1 Cognitive Structure

Our understanding of the brain, when appropriately interpreted and translated, can be richly informative of our understanding of the mind (Jack et al. 2006; Mather et al. 2013). For example, Neuropsychology, (the study of developmental and neurological deficits) reveals mental structure by identifying functional dissociations between cognitive processes (Shallice 1988). Brain imaging further extends this work, both by increasing our understanding of how cognitive processes map onto discrete neural regions, and by shedding light on the network organization of the brain. In particular, brain imaging methods have helped reveal the extent to which certain brain regions (and their associated cognitive processes) share functionally independent, cooperative and/or antagonistic relationships (Bressler and Menon 2010). Better understanding these relationships is significant to the current essay because it reveals the presence of a tension between different brain regions (Fox et al. 2005; Shulman et al. 1997b), and hence between their associated cognitive processes (Jack et al. 2012). In short, our neural structure imposes constraints on cognition, and hence on how we are able to make sense of the world. Behavioral measures and manipulations also inform our understanding of the relationship between cognitive processes, both within individuals and by examining differences between individuals.

Neuropsychology has demonstrated dissociations between different broad categories of cognition by examining the psychological profile associated with different mental disorders. Williams Syndrome is characterized by deficits in visuo-spatial reasoning, intact abilities to interpersonally connect with others, and a high degree of interpersonal warmth (for a review see Bellugi et al. 2000). At the other extreme, Psychopathy is characterized by social disconnection and pathologically low levels of interpersonal concern for others, but spared or even heightened abilities in analytic reasoning and in instrumental forms of social cognition, such as Machiavellian thinking and theory of mind (Babiak and Hare 2006; Blair et al. 1996). Autism spectrum disorder (ASD) is characterized by deficits in theory of mind skills coupled with spared feelings of concern for others (Blair 2005; Dziobek et al. 2008; Gleichgerrcht et al. 2013; Lockwood et al. 2013) and spared or augmented analytic cognition, including numerical, visuo-spatial and mechanical reasoning (Baron-Cohen 2002). These (double) dissociations between individuals with canonical presentations of each of these disorders provide evidence that the brain is divided into different systems for understanding the world.

Neuropsychological profiles also provide some indication of trade-offs between different types of cognition, such that individuals with deficits in one area may be particularly strong in others. For example, individuals with Williams often appear

unusually socially warm, some individuals with Psychopathy are unusually analytically intelligent and effective at interpersonal manipulation,[4] and some individuals with ASD demonstrate exceptional visuo-spatial abilities. This suggests the underlying cognitive processes may interfere with each other, such that a deficit in one cognitive domain 'releases' other cognitive domains, allowing these other cognitive domains to thrive. However, it is hard to draw this conclusion with confidence from Neuropsychological studies alone: First, it can be difficult to establish the appropriate control comparison group. Second, such trade-offs may be due to unusual developmental trajectories, such that individuals with a disorder tend to spend more time training up certain skills and neglecting others, even if the trained and neglected sets of cognitive skills might not naturally interfere with each other (e.g., individuals who are deaf tend to be better than controls at lip reading, however it is not believed that hearing interferes with lip reading). Additional empirical work is required to confidently infer that these systems of understanding interfere with each other.

Work in neuroimaging has provided additional evidence about brain regions associated with these different sorts of understanding. It has been noted for some time that various sorts of non-social cognition tasks, including mathematical, logical, mechanical, causal and visuo-spatial reasoning tasks, as well as focused attention and non-social working memory tasks, all tend to activate an overlapping set of brain regions (Corbetta et al. 1998; Duncan and Owen 2000; Fischer et al. 2016; Goel 2007; Jack et al. 2012; Martin and Weisberg 2003; Shulman et al. 1997b). Similar brain regions are also implicated in general intelligence, which correlates with non-social working memory performance (Toplak et al. 2011; Van Overwalle 2011) and is indexed well by visuo-spatial reasoning tasks such as Raven's matrices (Prabhakaran et al. 1997). These are the sorts of cognitive processes that tend to be compromised in Williams Syndrome. Initially, this network was often characterized as being responsible for reasoning in general (e.g., Duncan and Owen 2000). Further, following the functional division put forward by psychological dual process theory (Kahneman 2011; Tversky and Kahneman 1973), the supposed 'general reasoning' function of this network was often contrasted with other brain regions which were assumed to be involved in less evolved forms of automatic, heuristic or emotion-driven cognition (e.g., Greene 2007; Greene et al. 2004; Lieberman 2007).

However, it is now clear that this characterization is incorrect. First, recent evidence indicates that, while psychological dual process theory captures a valid division between more controlled vs. more automatic modes of processing, these modes generally occur within the same neural systems, rather than occurring in anatomically distinct neural systems (Dehaene and Cohen 2007; Evans and Stanovich 2013; Molenberghs et al. 2016; Van Overwalle and Vandekerckhove 2013). Second, a separate network of brain regions, sometimes referred to as the 'social brain', has been shown to be engaged by various types of social and ethical reasoning, as well as reflective, deliberative or self-regulatory forms of emotional cognition (Amodio and Frith 2006; Bzdok et al. 2012; Denny et al. 2012; Jack et al. 2012; Reniers et al. 2012; Schilbach et al. 2008; Van Overwalle 2009, 2011). These brain regions are also sensitive to cognitive load (Spunt and Lieberman 2014) and have their own social

[4] A fascinating extended discussion of the potential cognitive advantages associated with psychopathic personality traits can be found in Kevin Dutton's popular book 'The Wisdom of Psychopaths' (Dutton 2012).

working memory system that is anatomically distinct from non-social working memory systems (Meyer et al. 2015). These findings are inconsistent with dual-process theorists claiming that reflective and controlled "thinking engages a singular central working memory resource" (Evans and Stanovich 2013, p.226).

Hence, the first network does not, in fact, 'corner the market' on reasoning, nor is it responsible for all forms of deliberative, controlled or reflective cognition. Rather, just as the Neuropsychology of ASD and Psychopathy suggests, brain imaging also indicates we have distinct capacities for making sense of social and non-social aspects of the world. A striking demonstration of this division comes from a recent study (2016) which generated a semantic atlas of the brain while participants listened to several hours of natural narrative stories (selections from The Moth Radio Hour). Using data-driven methods, Huth et al. (2016) demonstrated that the brain divides very clearly between social/emotional concepts (e.g., 'communal', 'mental', 'social', 'emotional') and perceptual/analytic concepts (e.g., 'visual', 'tactile', 'abstract', 'numeric').

As neuroimaging has progressed, it has also moved beyond the mapping of cognitive functions to reveal relationships between brain areas and networks (for reviews see Bressler and Menon 2010; Fox and Raichle 2007; Sporns 2014). The most striking and widely cited such observation relates to the two networks mentioned in the introduction, the 'Task Positive Network' (TPN) and 'Default Mode Network' (DMN). Initially, researchers observed that a wide variety of tasks tended to activate the TPN (Corbetta et al. 1998; Shulman et al. 1997a), and that at the same time the same tasks tended to suppress the DMN below resting levels (Shulman et al. 1997b). The tendency for the DMN to be more active when participants were 'doing nothing' than when they were given a task was noted by many laboratories, and created considerable confusion and skepticism because the phenomenon violated the standard methodological assumptions of cognitive neuroscience (Price and Friston 1997). As a result, there was much speculative theorizing, in particular with regard to the function of the DMN. It was then observed that the 'see-saw' relationship between these networks was also present even when participants were not given any task (Fox et al. 2005), and is even present to a lesser degree in anaesthetized new-world monkeys (Vincent et al. 2007). These observations were significant because they suggested that the tension between these brain areas is not driven by task demands, but rather emerges from the network architecture of the brain. This suggestion has been supported by subsequent neural modeling work (Sporns 2014).

However, the cognitive significance of the tension between these networks remained hotly debated. Some researchers suggested the tension may occur because the DMN is involved in 'spontaneous cognition' or 'mind-wandering' (Mason et al. 2007; Raichle et al. 2001). Others (Andrews-Hanna 2011; Buckner and Carroll 2007; Buckner et al. 2008) suggested that the tension reflected competition between two sorts of attention – attention to the external world (e.g. visual stimuli) and attention to internal stimuli (e.g. memories). However, on closer examination, it became clear that the TPN mapped closely onto the analytic reasoning areas reviewed above, whereas the DMN mapped onto the social brain (Reniers et al. 2012; Schilbach et al. 2008; Spreng 2012).

Hence, inspired by our prior philosophical work concerning the possible cognitive origins of the problem of consciousness (Robbins and Jack 2006) we put forward another possible explanation. The opposing-domains hypothesis holds that the tension between the TPN and DMN reflects a cognitive tension between analytic and

empathetic thinking. We conducted an experiment which tested this hypothesis against the alternative hypotheses (spontaneous vs. task-oriented cognition, and internal vs. external attention). The findings (Jack et al. 2012) both ruled out the alternative hypotheses and provided strong support for the opposing-domains hypothesis. The social tasks used in this study activated all the key regions of the DMN well above resting levels, and also suppressed the key regions of the TPN below resting levels. This demonstrated, for the first time, that the TPN-DMN see-saw could be pushed to the opposite extreme (compared to analytic reasoning) by engaging social narratives. In addition, we observed that the brain regions which were pushed up and down by the scientific reasoning and social narrative tasks we gave participants mapped exactly onto the regions which display an endogenous alternating pattern in the absence of any task (i.e. at rest). This confirmed we had identified just the types of cognition that push the brain's internal see-saw. This is a finding we have replicated numerous times in our laboratory, using different tasks that press on the same processes. Although other researchers still claim the tension reflects internal vs. external attention, no other laboratories, either before or after, have been able to achieve the same key result using those or other cognitive processes.

The idea of a division, or tension, between analytic thinking on the one hand and social cognition on the other, has also been suggested by others. Paul Bloom (2004) makes remarks suggestive of such a tension in his book Descartes' Baby. However, his experimental work only indicates a dissociation between thinking about minds and objects – they do not provide clear evidence for a competitive relationship. Simon Baron-Cohen's empathizing-systemizing theory similarly suggests a trade-off between social and analytic reasoning (Auyeung et al. 2009; Baron-Cohen 2002; Baron-Cohen et al. 2001a). However Baron-Cohen's theory has been criticized (Andrew et al. 2008; Jack et al. 2016) because he does not clearly distinguish between different types of 'empathy', in particular between more cognitive social skills, such as theory of mind performance, and empathic concern for others – the latter of which is impacted more severely in Psychopathy than ASD.

In contrast to the finding that prolonged and engaging social/emotional tasks activate the DMN above rest (Hyatt et al. 2015; Iacoboni et al. 2004; Jack et al. 2012, 2013; Spunt et al. 2015) and deactivate the TPN below rest (Jack et al. 2012, 2013), a review of the literature demonstrates that more cognitive and performance oriented forms of social reasoning, such as theory of mind tasks, tend to activate both the DMN and TPN. These include decoding others' intentions (Schurz et al. 2014) and keeping social information in mind for performance related tasks (Meyer et al. 2015). These tasks appear to involve a coalition of neural resources from the DMN and TPN, which breaks with the usual tendency for networks to suppress each other.

We suggest that this coalition between the networks is cognitively limited in an important way, such that it involves only a shallow, distanced or disengaged, appreciation of other's (and one's own) mind(s). This claim is supported by the observation that co-activation of the TPN and DMN is associated with a variety of forms of instrumental and/or anti-social cognition, including: stigmatizing attitudes (Krendl et al. 2006, 2009), lying and deception (Christ et al. 2009), breaking promises (Baumgartner et al. 2009), dehumanization (Jack et al. 2013) and Machiavellian tendencies (Bagozzi et al. 2013). Hence, there is some evidence that engaging the TPN interferes with empathy. Similarly, there is excellent evidence that engaging the DMN interferes with the functions of the TPN, such

that activity in the DMN is associated with mind wandering, slips of attention and performance errors when participants are engaged in analytic reasoning tasks (for a review see Anticevic et al. 2012).

Behavioral studies also provide some evidence for a trade-off between analytic thinking and empathy. There is now converging evidence that inducing 'analytic' or 'calculated' mindsets, which are associated with TPN activation and DMN suppression, both increases antisocial tendencies and decreases prosocial sentiments (Rand et al. 2012; Small et al. 2007; Wang et al. 2014; Zhong 2011). For instance, solving quantitative GRE questions before performance-based games (e.g., ultimatum) explicitly increases the tendency to think analytically and behave deceptively and selfishly, while also reducing the tendency to think socially (Wang et al. 2014; see also, Zhong 2011). Manipulating participants to think about others as mere numerical objects (i.e., statistics), instead of identifiable humans, reduces sympathetic behavior in the form of charitable donations (Small et al. 2007). Similar work in neuroscience has revealed robust deactivation in key nodes of the DMN when participants perceive individuals to whom they have assigned economic value (Harris et al. 2014).

We suggest that pushing people into these sorts of analytic and performance-based mindsets, and thereby engaging the TPN, renders individuals temporarily insensitive to ethical sentiments, which are associated with the DMN.[5] This is different from merely selectively withholding or suspending certain moral sentiments towards others, or oneself (Bandura 1999). Instead, it involves actively engaging another sort of cognition, namely analytic cognition. In virtue of invoking analytic cognitive processes, social and moral dilemmas are stripped of key aspects of their 'social' and 'moral' character such that the dilemma is reframed as a strategic issue which is understood in 'cold' analytic terms. Indeed, when analytic cognition is fully engaged, the situation may no longer be perceived as an ethical *dilemma* at all, because the individual is rendered insensitive to ethical sentiments. Under these conditions, persons are not viewed as moral patients ('ends in themselves') but rather as physical objects or instrumental tools ('means to an end'). This tendency can be lessened by conditions which decrease activity in the TPN and increase DMN activation. The simplest example of such a condition is merely asking people to close their eyes, which disengages TPN regions associated with visuo-spatial attention and increases activity in the DMN (Nakano et al. 2013). Caruso and Gino (2011) show this simple manipulation heightens the perception of moral distinctions and discourages dishonest behavior.[6] Amit

[5] Dual-process theorists tend to conflate deliberate and reflective thought with analytical reasoning. For example, according to Rand et al. (2012), rapid Type 1 thinking is prosocial whereas deliberative thinking is less prosocial. However, this mapping makes it very hard for dual-process theorists to account for other findings that deliberate thought can increase ethical behavior, when compared to time pressured conditions (Gunia et al. 2012). Dual process theorists overlook the possibility that there are two sorts of reflective and deliberate modes of understanding, as articulated here. We have further advanced our account elsewhere to explain how the distinction between Type I and Type II thinking can be seen as orthogonal to the distinction between analytic and empathic thinking. When both orthogonal factors are taken into account, this allows for an explanation of such discrepant findings, and also generates further falsifiable predictions about the cognitive processes involved in (un)ethical behavior (Friedman et al. 2015).

[6] It is worth mentioning that Caruso and Gino's (2011) experimental design was not motivated by the observed tension between the TPN and DMN. That is, the authors did not reason that closing one's eyes might lead to increased ethical sensitivity by engaging participants' DMN and suppressing their TPN. Hence, in the general discussion, the authors candidly write that "our results do not provide a clear explanation for why closing one's eyes causes increased simulation [sensitivity to moral emotions] in the first place" (p.284).

and Greene (2012) show the same manipulation biases people against utilitarian judgments and towards deontological judgments.[7]

Behavioral measures can also index individual differences in the tendency to adopt the cognitive processes associated with the two networks (see Figure 1a). For instance, self-reported empathic concern, as measured by the Interpersonal Reactivity Index - Empathic Concern (IRI-EC; Davis 1983) – or the absence of its converse, callous affect – (Self-Report Psychopathy, Callous Affect; SRP-CA (Paulhus et al. 2009) – is the signature other-oriented emotion characteristic of engaging the DMN and deactivating the TPN. These measures also index the primary personality characteristic of Psychopathy. Measures of analytic reasoning such as Shane Fredericks Cognitive Reflection Test (CRT) and the Intuitive Physics Test, (IPT) (Baron-Cohen et al. 2001b) are believed to index a tendency for sustained engagement of the TPN and deactivation of the DMN. Finally, there are a number of established measures for assessing the ability of individuals to engage in instrumental social reasoning, the domain most affected by ASD and which is believed to require a coalition of cognitive processes with substrates in both the TPN and DMN (e.g., Reading the Mind in the Eyes; Baron-Cohen et al. 2001b; Interpersonal Reactivity Index-Perspective Taking [IRI-PT], Davis 1983; Diagnostic Analysis of Nonverbal; Accuracy [DANVA], Nowicki and Duke 2001). These measures of individual difference also provide some evidence for a trade-off between the functions of the TPN and DMN.

In a series of five studies, we found evidence for a modest ($r \sim -0.2$) negative correlation between the empathy and analytic thinking measures cited above (although it is not the primary topic of the paper, some of these findings are reported in Jack et al. 2016). We also found associations with gender, such that females scored slightly lower on analytic thinking measures and slightly higher on empathy measures, however the trade-off was present even within gender categories. Measures of instrumental social cognition showed no negative correlations, and positive associations were greater with analytic thinking measures than with empathy measures.

In conclusion, evidence from numerous fields of research in cognitive science suggests a tension between different types of cognition which guide quite different facets of uniquely human understanding. On the one hand, we possess a remarkable ability to understand and appreciate the minds of others – a type of intersubjective understanding which generates a sense of interpersonal connection and the feeling we have been 'understood' by others. On the other hand, we possess a highly sophisticated ability to represent and manipulate the physical world and to make sense of abstract rule-bound systems (e.g., logic, mathematics). The existence of a tension between these two types of understanding is indicated by more nuanced Neuropsychological profiles (i.e., profiles which do not merely focus on deficits) and by behavioral studies. However, this tension became particularly salient when imaging technology advanced to a point where it allowed us to observe the brain in action. In cognitive neuroscience, the pronounced tendency for a trade-off in activity between two large scale cortical networks was noted, and went on to generate widespread interest and debate, some fifteen years before its cognitive significance became well understood.

[7] Amit and Greene (2012) explain their results by combining dual-process theory with construal level theory. We believe the current framework provides a simpler and more comprehensive account of their findings.

The tension between these two styles of thinking raises an important question, which is unfortunately very hard to answer with confidence: Why has our neural architecture evolved to create this tension? Given evidence that these two types of thinking often interfere with each other, that high IQ is associated with greater tension between the networks (Anticevic et al. 2012), and that many mental disorders are marked by decreased tension between the networks (Broyd et al. 2009; Buckner et al. 2008), we suggest it was a matter of evolution selecting for the most cognitively efficient design. Broadly speaking, 'analytic' cognition appears best suited to representing phenomena which are physical/inanimate, or abstract systems well defined by rules, and uses a focused and systematic approach to generate predictions. In contrast, 'empathic' cognition appears best suited to representing phenomena using subjective concepts (e.g., emotions) that are vague in definition, and whose application is highly context dependent. Such phenomena appear best understood by a broader focused, more holistic or 'synthetic' form of reasoning. If this characterization is broadly correct, such that quite different cognitive processes evolved for the purposes of allowing us to understand and engage with minds vs. objects, then it is not surprising that our neural architecture also evolved to help avoid these different types of thinking from interfering with each other.

We suggest the relationship between these two neural networks (DMN & TPN), and the distinct types of cognition they instantiate (empathic and analytic), presents an endogenous barrier to successfully integrating the concepts and representations particular to each of their opposing cognitive domains. In a very real sense, the way in which certain phenomena are understood or represented by one domain transcends explicability in terms of the other domain. This gives rise to incommensurable worldviews not just between individuals, but even *within* a single individual. That is to say that we have a divided brain, and we cannot transcend these endogenous constraints by subsuming disparate viewpoints and their related concepts under a more unified and ultimately transcendent mode of understanding. It appears, whether it is regarded unfortunate or not, that evolution has simply not equipped us with any such transcendent faculty of understanding. Put another way, it appears nature has selected to promote in us the mundane skills required to effectively navigate our social and non-social environments, and has rudely ignored the nobler goal of promoting our ability to generate a parsimonious metaphysics. In the next section, we expand further on the details of how this unfortunate fact might explain some persistent debates in philosophy.

2.2 Schisms in Understanding

In this section, we transition between our 'cognitive scientist hats' and 'philosopher hats' in order to articulate the relevance that the antagonistic relationship between the DMN and TPN bears to philosophy. As we mentioned in the introduction, this is facilitated by adopting an intermediate level of description between psychology and philosophy. We thus refer to three cognitive stances one can adopt in order to make sense of the world and one's experience of it in general, and certain philosophical problems in particular (Robbins and Jack 2006). The Physical stance, which is adopted in such disciples as physics, chemistry, engineering and biology, represents the physical structure of material objects and makes predictions based on physical laws. The

Intentional stance, which is adopted, for example, in economics, represents the beliefs and desires of agents and makes predictions based on the assumption they are rational. The Phenomenal stance is not aimed at prediction, but represents the experiential states of persons as well as generating moral sentiments (i.e., represents others as moral patients and moral agents).

As can be seen in Fig. 1, we have mapped each of these stances onto different configurations between the TPN and DMN. The Physical stance maps onto TPN activation and DMN suppression; the Phenomenal stance onto DMN activation and TPN suppression; and the Intentional stance onto co-activation of both networks. As a reminder, we are not claiming that these stances exhaust the cognitive processes associated with each of these networks or their more nuanced configurations that future research might reveal. Instead, we are claiming that each of these stances recruits aspects of the broader types of understanding associated with each neural signature. Consequently, each stance inclines an individual to appreciate certain philosophical considerations from its own vantage point, while ignoring or misinterpreting considerations that emerge from other vantage points offered by other stances. The antagonistic relationship between the networks underlying these stances prevents us from integrating all of their unique considerations into a single, coherent stance or 'solution'. This is why we have mapped each stance onto opposing views related to our three problems of interest.

According to our original account (Robbins and Jack 2006), the Physical and Phenomenal stances are in fundamental tension, whereas the Intentional stance has both some tension and some interplay with the Physical and Phenomenal stances. Translating this framework to traditional issues in philosophy, Robbins and Jack (2006) claimed that the fundamental tension between the Physical and Phenomenal stances means that attempts to translate between these two stances suffers from a deep and salient 'explanatory gap' (Levine 2000). We regard the mind-body problem, and more specifically the problem of consciousness as famously treated by Nagel (see below), as emerging from the tension between these stances. Whereas the Physical stance inclines the individual to consider the physical system and mechanistic behavior of neural processes associated with mental states, the Phenomenal stance inclines the individual to appreciate the experiential nature of our own, and others, mental states. Of the three problems discussed here, we will argue that this is the most intractable precisely because the tension between these cognitive stances is the most pronounced (see Figure 1).

Robbins and Jack (2006) further point out that Brentano's problem[8] is associated with a somewhat less salient and deep, but still present, tension between the Physical and Intentional Stances (see Figure 1D). Here, we are adding to this scheme (see below) the observation that Strawson's (1962) work appears to astutely identify a point of tension

[8] Brentano's problem concerns the intentionality of mental states, i.e. that beliefs and desires are *directed at* or *about* something. The problem that Brentano is often credited with raising is the apparent difficulty of accounting for the intentionality of mental states by appeal to the purely physical properties of mental states (see Haldane 1989 for a more nuanced account of Brentano's views). In other words, one might say that Brentano is known for pointing out an apparent explanatory gap between the physical and the intentional. Most philosophers regard this apparent explanatory gap as distinct from the 'hard' problem of consciousness (i.e. the apparent gap between the phenomenal and the physical). However, Rosenthal's (2005), higher-order-thought theory of consciousness views the gap between the phenomenal and physical as either subordinate to the gap between the intentional and the physical, or simply identical to it.

between the Intentional and Phenomenal stances (see Figure 1e). More specifically, the claim is that the opposition between compatibilism and incompatibilism reflects an inability to effectively unify the considerations about human minds that are afforded by these two stances. Whereas the Intentional stance emphasizes an entity's agentic capacities, primarily for one's own predictive or explanatory purposes, the Phenomenal stance emphasizes an entity's capacity for experience and their moral patiency.

Lastly, we have suggested (Jack et al. 2014; Rochford et al. 2016) that the historical tension between utilitarian and deontological ethics emerges from the observation that each system emphasizes considerations that are more characteristic of blending the Physical and Intentional stances (utilitarianism) or the Phenomenal and Intentional stances (deontology). Considerations that are central to each ethical system, but appeal to these different blends of the relevant stances, prevents us from understanding one ethical system in terms of the stances associated with the other. In other words, key tenets of each system are lost in translation when cycling between the three cognitive stances.

Because each of these stances are associated with different psychological processes, and well suited measures to assess individual differences in these processes (section 2.1) the foregoing claims are empirically falsifiable hypotheses that are susceptible to the methods of cognitive science. As we discuss throughout the essay, this cognitive mapping allows for testing hypotheses designed to evaluate how the tendency to adopt each of these stances relates to one or another position pertaining to the three issues discussed here. It is obviously also possible to test these claims using other behavioral methods and the methods of functional magnetic resonance imaging (fMRI). In the case of moral decision making, neurological findings reported by Joshua Greene and colleagues (Greene et al. 2001, 2004) map extremely well onto our framework (Jack et al. 2014),[9] and we have also extended his work in the moral domain with further work on the perception of moral patiency and dehumanizing.

Before we close this section, we wish to address two points which have generated some confusion in the past. First, we have claimed that the Intentional Stance represents a co-activation of regions within the TPN or DMN, or more specifically a coalition of neural resources from both networks. As discussed above, this claim is based on evidence that tasks (e.g. theory of mind tasks) which rely on the Intentional Stance have this neural signature. As discussed above, the tendency for the TPN and DMN to be antagonistic, or demonstrate a see-saw relationship, is a clear and striking observation, which can be seen both during the performance of a wide variety of tasks and during the resting state. However, like most biological phenomena, the see-saw relationship between these networks is neither rigid nor absolute. There is evidence that the see-saw relationship is 'broken' during tasks involving instrumental social reasoning (see above), at moments of creative insight (Beaty et al. 2014), and that it is weaker in a variety of mental disorders (see above).

A metaphor which may be helpful is to think of the TPN and DMN as similar to two opposing political parties. In general, they adopt different positions, and there is considerably better cooperation between members of the same party than between members of opposing parties. Nonetheless, bipartisan coalitions can be formed which

[9] Note that Joshua Greene interpreted his findings through the lens of psychological dual process theory. As discussed in the previous section, the evidence is now clear that this theory does not provide an adequate account of the tension between the TPN and DMN. We have provided more consistent interpretations of such work elsewhere (Jack et al. 2014; Rochford et al. 2016).

can be both productive and efficient. Such coalitions are not the norm, and they are often unstable. Further, while the positions such bipartisan coalitions adopt are usually closer to the positions of either party than the two parties are to each other, and borrow to some degree from each one's ideology, they also usually demonstrate some tension with the positions of each of the parties in isolation.

Second, we have claimed that both utilitarianism and deontological systems of ethics can be best understood as blends of stances (Utilitarian - Physical and Intentional; Deontological - Phenomenal and Intentional). To be clear, it is not our claim that we have no ability to translate between stances. Our claim is that the concepts that underlie different stances are incommensurable. It is nonetheless possible for us to weigh considerations that derive from different modes of understanding, and indeed our view is that ethical judgments are best accomplished by balancing between the considerations generated by different perspectives, i.e. between utilitarian considerations on the one hand, and deontological considerations on the other (Jack et al. 2014; Rochford et al. 2016). Further, since both utilitarian ethics and deontological ethics reflect systems of thought, rather than the more specific conceptual gaps which drive the problems of consciousness and free will, it is perhaps unsurprising that these systems of thought already achieve some integration across different stances.

Our approach to all three of these philosophical issues, of consciousness, free will, and ethics, is Kantian in flavor in the sense that we deliberately put off the question of which of two incompatible philosophical positions is correct, pending a better understanding of why the disagreement occurs (see introductory quote and section 3.4). By looking first to understand the cognitive mechanisms engendering different responses to these problems, we seek to address the question of whether these responses are truly contradictory, or instead merely represent distinct modes of apprehension and understanding. In doing so, our approach draws on empirical evidence and philosophical considerations in an attempt to illuminate both the origin *of* schisms related to perennial problems and, further, the perception *that* such problems represent more than mere errors of philosophical reasoning or scientific ignorance.[10]

[10] When it comes to perennial philosophical problems, we regard the 'origin *of*' the problems and the 'perception *that*' the problems are intractable, as distinct but related phenomena. Here is why: Some philosophers argue that clarifying the origins of a particular perennial problem will render it unproblematic. For example, that a mature neuroscience will dissolve the 'Hard' problem of consciousness (Churchland 1981; Dennett 1991, 2013). Philosophers espousing this view indeed acknowledge that there is a (temporarily) genuine and (temporarily) unavoidable 'origin of' the problem – namely, scientific ignorance. Hence, a mature neuroscience should dissolve the 'perception *that*' the problem is intractable. Time and scientific progress are the obstacles, not endogenous neural constraints. Others have claimed that philosophers and scientists are actually wasting their time trying to solve such problems as the 'Hard' problem of consciousness because it is nothing more than an intentionally designed philosophical fiction (Machery 2013; Sytsma and Machery 2009). Although Machery (2017)has since updated his perspective on this and other metaphysical issues (see section 4), anyone else who subscribes to this sort of view denies that there is any genuine 'origin *of*' the problem, and consequently argue that the 'perception *that*' it is intractable is a non-sequitur. On this view, the philosophically naïve shouldn't perceive the problem because they haven't been exposed to philosophical discourse. The origin of the explanatory gap, for example, is traced to modern philosophers (e.g., Descartes 1641/1996; Leibniz and Montgomery 2005), not our endogenous neural constraints. The foregoing treatments do not practice the skeptical method. They are deflationary accounts of what we, like Kant (and perhaps James, Nagel and Strawson), view as deeply meaningful problems that are bound to persist despite advances in scientific and philosophical understanding.

3 The Illumination of Conceptual Tension

In section 2.1 we outlined a model of cognitive structure based on research in the cognitive sciences. In section 2.2 we then identified three philosophically relevant cognitive stances that map onto this model of cognitive structure, and we outlined how a number of long running philosophical debates can be seen as reflecting tensions between these stances. In this section, we highlight the similarities between this empirical analysis of these debates and examples of a type of philosophical analysis we call 'the illumination of conceptual tension'. This approach is exemplified by influential philosophical contributions by Immanuel Kant (Kant 1787/1998), William James (James 1906/1975), Peter F. Strawson (1962), and Thomas Nagel (1974).[11] Notably, these contributions did not follow the usual models for conceptual analysis, such as developing a theory of the concept *x* or determining whether concept *x* was accurately deployed. Instead, they sought to illuminate the relationship between disparate and even opposing concepts or attitudinal viewpoints which, when used to represent the same issue, gave rise to opposing viewpoints about the issue. In doing so, they each attempted to give a descriptive account of what might be going on in us – whether at the conceptual or proto-psychological[12] level – that might explain our inability to reconcile the opposing viewpoints.

3.1 Nagel and the Problem of Consciousness

In *What is it like to be a bat?*, Nagel clarified the intractable nature of the mind-body problem by illuminating a disparity between two different viewpoints – one we adopt when representing our own subjective experience and the other we adopt when representing an objective and mechanistic explanation of a phenomenon. Nagel not only clarified the presence of a tension between these viewpoints, but articulated its relevance to the impossibility of ever knowing what it is like to be a bat *for* a bat, namely because objective explanations culminate in a complete removal from the subjective viewpoint (i.e., the further away science moves from a particular viewpoint, the better the scientific explanation). Such a disparity between perspectives could, Nagel suggested,[13] explain why even if psychophysical identity statements were necessarily true, they could still be conceived as contingent. This is because Nagel entertained the possibility that the conceptual frameworks for representing subjective

[11] Carnap (1955, 1962) also championed a similar type of analysis, although he did not apply it to the problems we are interested in here. Shepherd and Justus (2015) have lucidly explained how synthesizing the methods of experimental philosophy with "Carnapian explication" can have philosophical import in a parallel way to conceptual analysis more broadly. While we focus on different philosophical issues than Shepherd and Justus (2015), and do so through the lens of psychology and neuroscience, both projects take a more inclusive approach to experimental philosophy by arguing that it can have significant philosophical import that complements and extends historically important forms of philosophical analysis.

[12] We take it that reference to different 'faculties' for understanding, 'differences in temperament' and 'types of imagination' are best understood as claims about human cognition. However, such claims were not embedded in a well-developed scientific psychology and it was unclear how to test them. Hence we use the term 'proto-psychological'.

[13] In footnote 11 of *What it is like to be a bat?* Nagel discussed how 'different types of imagination' might give rise to a division between the sort of understanding we use for understanding brain states on the one hand, and subjective experiences on the other.

experience and physical mechanism are not only independent, but they are so disparate that we could not even begin to imagine how they refer to the same thing (e.g., that c-fiber firing just *is* the immediate phenomenological experience of pain). Indeed, one conceptual framework treats the phenomenon in question as existing *from* a viewpoint, and thereby recruits distinct conceptual components for fulfilling that representation, while the other treats the phenomenon *as* a mechanistic process that is neutral to any particular viewpoint – 'the view from nowhere' – which in turn has its own conceptual components. This engenders a seeming incommensurability. As Nagel wrote:

> "I believe it is precisely this apparent clarity of the word "is" that is deceptive. Usually, when we are told that X is Y we know how it is supposed to be true, but that depends on a conceptual or theoretical background and is not conveyed by the "is" alone. We know how both "X" and "Y" refer, and the kinds of things to which they refer, and we have a rough idea how the two referential paths might converge on a single thing, be it an object, a person, a process, an event, or whatever. But when the two terms of the identification are very disparate it may not be so clear how it could be true. We may not have even a rough idea of how the two referential paths could converge, or what kind of things they might converge on, and a theoretical framework may have to be supplied to enable us to understand this. Without the framework, an air of mysticism surrounds the identification" (1974, p. 447).

Nagel's goal was primarily descriptive, so he neither offered a solution to the mind-body problem nor proposed how we might build the framework for explicating the link between the seemingly incommensurable subjective and objective modes of representation. That this framework seems an impossible one to build, at least to some, is, we suggest, not due to scientific or technological ignorance. Instead, we suggest it is because what Nagel described as the subjective view point and the 'view from nowhere' can be mapped onto the Phenomenal and Physical stances, respectively. Whereas the Phenomenal stance inclines us to appreciate the experiential nature of 'what it is like to be' in a mental state, the Physical stance inclines us to reduce a mental state to its neurobiological mechanisms.

We suggest the neurological tension underlying these two incommensurable modes of understanding helps explain why there *is not* an 'air of mysticism' when we reduce our concept of a physical thing, such as water, to its more basic physical constituents, namely H_2O. This is because no cross-domain conceptual mapping is required by this identity, but just a reduction of one physical concept to another physical concept. In other words, all of the conceptual reduction in the identity statement "water is H_2O" is completed within the same system, that is, by adopting the Physical stance and engaging its underlying neural hardware. This is not the case when attempting to move from the Physical stance to the Phenomenal stance, or vice versa.

In contrast to Kim (1984, 2007), our view is that an attempt to 'reduce' the phenomenal character of a mental state into its physical constituents should not be seen as a reduction from one level of explanation to a lower one, but instead as a type of conceptual translation between fundamentally distinct explanatory frameworks. As Nagel (1974) so vividly articulates, and Chalmers (1997) repeats, it just *seems* that

something is left out from this translation. The account outlined here provides evidence that we are neurologically constrained from fully translating such disparate concepts. In a very real sense, the way in which certain phenomena are understood or represented by one domain (e.g., the Phenomenal stance and its underlying neural signature) transcends explicability in terms of the other domain (e.g., the Physical stance and its underlying neural signature). The harmonious communication required for successfully translating between these domains is not fully realizable (see Figure 1).

In support of this is behavioral evidence that a tendency to adopt the Phenomenal stance, assessed by individual differences in empathic concern (IRI-EC; see section 2.2) is associated with belief in dualism (Jack 2014) but unrelated to measures of analytic (assessed by the CRT) and physical reasoning (assessed by the IPT) ability (Jack et al. under revision). Both of these measures are associated with TPN activation and DMN suppression, while the latter is characteristic of the Physical stance in particular. We suggest that a natural inclination to adopt the Phenomenal stance draws attention to what Nagel described as the 'what it is like to be' aspect of consciousness, which in turn inclines an individual to reject purely mechanistic and physicalist accounts of consciousness because they seem to leave something out – namely subjective experience. A common strategy to 'fix' the feeling that 'something is left out' is to posit belief in dualism. Individuals who lack empathy and experience difficulty in *understanding* the 'what it is like to be' aspect are less likely to perceive the gap, and thus more inclined to adopt a physicalist view of the mind. Hence, we suggest the tension between the DMN and TPN partially explains the presence of the explanatory gap (Levine 2000), and thus motivates belief in dualism – or in Nagel's terms, the apparent contingency of the relationship between brain states and mental states.

If Nagel had had access to this evidence in 1974 and privately used it to inform his treatment of the conceptual tension underlying the mind-body problem, would we think of his project as any less philosophical? Or, what is only slightly different, should we *now* treat Nagel's analysis differently in light of the emerging scientific evidence in favor of his general conclusion? This perspective seems nonsensical. The philosophical import of the emerging experimental evidence is both intriguingly analogous and complementary to the philosophical import of Nagel's original analysis. What Nagel did was articulate the presence of a conceptual tension emerging from different viewpoints that we impose on our experience of the world. He sought to capture the disconnect, *in us,* between different modes of understanding and conceptual representation, the nature of which poses an obstacle to unifying the concepts into a coherent idea.

One of the many reasons we admire Nagel's analysis is because it aimed to clarify "the point of misunderstanding in disputes that are honestly intended and conducted with intelligence by both sides" (Kant 1787/1998, A424/B452). In other words, Nagel was sympathetic to the disagreement among expert philosophers. He did not attempt to solve the problem or adjudicate between monistic and dualistic theories of consciousness, but instead articulated why these two views are opposed, that is, why the problem seems so persistently intractable. By grounding the problem in our cognitive architecture, this is precisely the goal to which our project seeks to contribute. For us, the problem (referred to either as the 'mind-body problem' or more specifically the 'problem of consciousness') itself becomes no less genuine, but certain goals that drive many philosophical and scientific approaches to the problem certainly look less

sensible. In particular, the common philosophical approach of attempting to arrive at a definitive solution in favor of one view over another looks futile. Second, the scientific hope that a reductive account of consciousness can be found looks like a 'category mistake' (Ryle 1949) – the mistake being to suppose that a theory which clarifies our mechanistic understanding of the brain from the perspective of the Physical Stance could also succeed in clarifying our experiential understanding of ourselves and others from the perspective of the Phenomenal Stance.

3.2 Strawson and the Problem of Free will

In *Freedom and Resentment,* Peter Strawson clarified that the incommensurable positions pertaining to the dilemma of determinism – to adopt compatibilism vs. incompatibilism – emerged from a tension between different conceptual and attitudinal viewpoints. He further argued that certain concepts and attitudes share a much stronger affinity with one viewpoint than the other (i.e., the participant viewpoint and its reactive and vicarious attitudes vs. the objective viewpoint and its association to a 'one-eyed utilitarianism' in which the only justification for punishment derives from appeals to social policy/control). The viewpoint one adopted influenced their stance on the problem. In treating the issue this way, Strawson argued that the truth or falsity of determinism is irrelevant to moral responsibility because our feelings of moral condemnation (e.g., resentment, gratitude), and the relevant concepts and social practices (e.g., justice, punishment), are not founded on a theoretical understanding of the way the world is or how others behave. Instead, they are founded on our deep and genuine commitment to the well-being of ourselves and others, that is, our morally reactive attitudes and their vicarious analogues.

Strawson clarified this by emphasizing that his 'optimist/compatibilist' divorces herself from our uniquely human moral sentiments by focusing on the social utility of punishment and control. Strawson's 'pessimist/incompatibilist' is aware that leaving out the morally reactive attitudes "excludes at the same time the essential elements in the concepts of *moral* condemnation and *moral* responsibility" (Strawson 1962, p.21) and, out of reverence for these sentiments, puts forth dubious metaphysical propositions (i.e., claims at odds with a naturalistic worldview; 'panicky metaphysics[14]', in Strawson's terms). But the metaphysical propositions only seem dubious because what Strawson referred to as the objective and participant viewpoints are not just distinct; they "are, profoundly, *opposed* to each other" (Strawson 1962, p.9). Indeed, in illuminating the tension between these opposing conceptual attitudes, Strawson noted that "But what is above all interesting is the tension there is, *in us,* between the participant attitude and the objective attitude. One is tempted to say: between our humanity and our intelligence." (Strawson 1962, p.10). Hence, Strawson emphasized how our cognitive faculties, and the relationship between them, might relate to different perspectives on this philosophical issue.

[14] Strawson used the term 'panicky metaphysics' in his essay *Freedom and Resentment* to refer to ideas and beliefs that claim a departure from a naturalistic or empirically driven account of actions in which humans are freely exercising their agency (e.g. the libertarian view of freedom, or 'contra-causal freedom'). Strawson describes such beliefs as 'panicky' because, he argues, they represent an intellectually inadequate knee-jerk defense against the fear that determinism undermines our humanity.

An interesting question is whether Strawson and Nagel were illuminating the same conceptual tension, a fundamentally different tension, or a related and overlapping conceptual tension. It is clear that both Strawson and Nagel contrast viewpoints which differ critically in their degree of 'objectivity' and 'distance' from the entity who is the presumed locus of experience and/or will. For Nagel, the objective viewpoint is a completely removed and reductionist scientific perspective, the 'view from nowhere'. For Strawson, the objective viewpoint is not totally divorced from the human perspective but is exemplified by the temporary suspension of particular "moral reactive attitudes" towards a particular agent. In short, it is a *morally* detached viewpoint that enables us to try to "understand[ing] how he [an agent] works" (Strawson 1962). This viewpoint has a close affinity with Dennett's Intentional stance, which is also part of our philosophical theory of cognition (see Figure 1). Indeed, for Strawson, the objective viewpoint of 'understanding how he works' is not an understanding aimed at the reductive/subpersonal neurological or physical level (i.e., the Physical stance) but is instead at the level of mental states (i.e., the Intentional stance). Hence, it appears that while Nagel illuminated a tension between the Phenomenal and Physical stances, Strawson illuminated a tension between the Phenomenal and Intentional stances. Our account holds that the intentional stance is realized by a blend of the cognitive capacities that underlie the physical and phenomenal stances (see Section 2 and Figure 1). Hence, we suggest that the conceptual tensions illuminated by Nagel and Strawson derive from the same neurological tension (i.e. between the TPN and DMN), even though the 'objective' viewpoints they reference correspond to distinct cognitive stances (i.e. the Physical and Intentional Stances).

There is further empirical support for Strawson's analysis that recourse to the "panicky metaphysics of libertarianism" (Strawson 1962, p.25) is driven by moral sentiments. More specifically, emerging evidence is beginning to suggest that it is a desire for retributive justice – not the utility of social control and deterrence – which is related to the tendency to "go beyond the [physical] facts" (Strawson 1962, p.20) and endorse libertarianism (Clark et al. 2014; Jack et al. under revision; Shariff et al. 2014). This behavioral work establishing a link between moral sentiments and "panicky metaphysics" not only supports Strawson's (1962) analysis, but also provides support for our (neurologically informed) account: Moral sentiments are, by hypothesis, tied to the Phenomenal Stance. When these sentiments are salient to the individual, the tension between the Phenomenal and Physical stances causes them to be drawn to metaphysical world views that depart from physicalism or naturalism. We suggest that deficits in these sentiments make the viewpoint associated with the Intentional stance more readily adoptable.

As with Nagel, we might ask: If Strawson somehow had access to this evidence in Strawson 1962 and privately used it to inform his treatment of the dilemma of determinism, should we regard his project as any less philosophical? Or, what is only slightly different, in light of the emerging evidence in favor of his general conclusion, should we now regard it as less philosophical? We disagree with Knobe (2016) that a novel metaphilosophical framework needs to be constructed to fully appreciate the philosophical import of evidence which contributes to the 'same sort of thing' as previous philosophically significant analyses. One of the reasons we admire Strawson's analysis is because he sought to illuminate "the point of misunderstanding in disputes that are honestly intended and conducted with intelligence by both sides" (Kant, Kant

1787/1998). That is, he was sympathetic to the disagreement among expert philosophers. By grounding the problem in our cognitive architecture, this is precisely the goal to which our project seeks to contribute. For us, the problem of the apparent incompatibilism between freewill and determinism is no less genuine, but the quest for a compatibilist solution certainly looks less sensible.

In sum, both Nagel and Strawson sought to illuminate the origin of conceptual tension and clarify *why* the different perspectives to these problems are irreconcilable.[15] In doing so, they appealed to certain concepts or faculties *in us* that might engender the opposing perspectives. We contend that this most closely resembles the philosophical contributions afforded by our approach to experimental philosophy and our work in cognitive science. The most salient difference between our work and theirs is the methodology and tools we use to arrive at and support our explanation, not the philosophical significance of the explanation.

Despite the fact that both Nagel and Strawson provided what are broadly viewed as highly philosophically illuminating treatments of the problems, the problems still remain. Their resistance to be integrated into a naturalized worldview is telling. As far as we know, Immanuel Kant was the first person to articulate a reason as to why that may be the case.

3.3 Kant's Antinomies

Kant's project in the *Critique of Pure Reason* (Kant 1787/1998) is often referred to as the Copernican revolution of cognition/philosophy. His focus was not on a mind-independent world, but on the subject and the *a priori* conditions (e.g., pure forms of intuition, categories, schemata, etc.…) without which experience of objects would be impossible. Despite their necessity, some of these conditions inevitably 'overstep their bounds' and engender certain illusions about the world (and ourselves).

One such illusion is an antinomy, which Kant characterized as a contradiction between mutually exclusive metaphysical theses that are each internally coherent and equally plausible. Kant's claim that their (im)plausibility is dependent on transcendental idealism or Kantian suppositions of space and time is irrelevant for our purposes. We draw parallels between Kant's work and our work for a number of reasons, none of which hinge upon these aspects of his philosophy.[16] Primarily, Kant was the first philosopher to advance an account that made claims about how our cognitive constitution might give rise to contradictions between different sorts of understanding, and

[15] Admittedly, Strawson did try to reconcile the two incompatible perspectives, but his attempt required 'radical modifications' by both parties. Similarly, Nagel (1974), Chalmers (1996) and others have alluded to potential strategies for reconciling the phenomenal and the physical. However, we would argue the major (and most widely agreed) contribution made by all these philosophers is that of helping to illuminate the nature of the conceptual tension: What is important for the present purposes is that Strawson's analysis clarified that each conceptual attitude either leaves out something crucial (compatibilism leaves out moral sentiments) or introduces something dubious (incompatibilism introduces the 'panicky metaphysics of libertarianism'). The question of what to do about the tension is not only contingent on the tension being correctly identified, but is also an additional step which may or may not be warranted or advisable.

[16] Our most sincere thanks to Chin-Tai Kim. His excellent undergraduate seminar on Kant, attended by J.P.F. while J.P.F. was an undergraduate researcher in A.I.J.'s lab, provided an opportunity for us to recognize the remarkable and surprising parallels between Kant's work and the direction the lab's cognitive neuroscience research was pointing.

the relationship that each shares to opposing philosophical positions. We suggest these aspects of Kant's philosophy can be considered a type of proto-psychology, insofar as he tried to construct an account that would explain how cognition of objects is possible. Yet, despite his emphasis on our cognitive constitution, no one questions the philosophical import of Kant's work (whether you agree or disagree with his philosophy).

We also appeal to Kant because of the two criteria by which he characterized his antinomies. One criterion is that the competing propositions should be universally recognized across humans. We do not 'choose' to confront or construct these paradoxes – they simply emerge from our cognitive constitution. The second criterion was that the origin and cause of the antinomy is "not merely an artificial illusion that disappears as soon as someone has insight into it, but rather a natural and unavoidable illusion, which even if one is no longer fooled by it, still deceives though it does not defraud and which thus can be rendered harmless but never destroyed….from this there must arise a contradiction that cannot be avoided no matter how one may try" (A422/B450).

We regard both the mind-body problem and the dilemma of determinism as just such illusions, the source of which can be traced to the tension between the DMN & TPN. These issues are perceived cross-culturally (Kant's first criterion) (Bloom 2004; Sarkissian et al. 2010) and seem to (partially) emerge from an evolved feature of our neurology (Kant's second criterion). And although not an 'illusion' itself, we believe this relationship is applicable to the historical debate between deontological and utilitarian ethics, which is discussed in the next section.

Unlike William James (see below), Kant didn't trace the origin of competing views comprising his antinomies to idiosyncratic aspects of the cognitive constitution of individuals espousing each view. Instead, he traced them to two fundamental and disparate modes of understanding that everyone possesses. In Kantian parlance, these are theoretical reason and practical reason (and the incommensurable 'Ideas' legislated by each). Whereas theoretical reason posits certain concepts or principles that are necessary for advancing our empirical understanding of the world, practical reason posits certain concepts or principles which inform moral action. Kant's antinomies emerged by failing to perceive and appreciate this distinction. He thus delineated the disparate domains to which these two sorts of understanding are 'for', and restricted the bounds of each *precisely in order to* preserve their unique affordances. For illustrative purposes, Kant noted that the theses of his antinomies, which were associated with practical reason, had a "certain practical interest" and served as the "cornerstones of morality and religion", while their antitheses, which were associated with theoretical reason, "robs us of all these supports, or at least seems to rob us of them" (Kant 1787/1998, A465/B493-A469/B497).

This tension that emerges between Kant's theoretical and practical reason, at least when they 'overstep their bounds' by trying to make sense of phenomena[17] for which they are not properly suited, broadly parallels the tension reflected between the TPN and DMN. Indeed, our view is that each network is properly suited for opposing domains of cognition – one for perceiving the physical world and analyzing its structure; the other for appreciating phenomenal experience and recognizing humanity.

[17] Phenomena in the contemporary sense of the term, not the Kantian sense.

3.4 The Deontology vs. Utilitarianism Debate

We suggest that the observed neural tension between the TPN and DMN can also provide some insight into the deontological vs. utilitarian debate, at least as it relates to certain types of moral judgments (Jack et al. 2014). We are not here to defend one ethical school over the other, either in general or viewed from the perspective of its most notable proponents. Instead, our aim is to provide a theoretically driven and empirically supported account that aims to clarify the cognitive processes associated with the tendency to endorse deontological or utilitarian judgments in a particular type of moral dilemma – hypothetical footbridge dilemmas that aim to pit these opposing ethical perspectives against each other.

Ethical thinking has long been dominated by these two opposing frameworks attempting to guide moral action. According to deontological thinking, which is famously exemplified by the work of Immanuel Kant, right moral actions emerge from a sense of moral duty. They are motivated by some abstract rule, such as the categorical imperative (i.e., never will an action that you yourself could not will to become a universal law), or some interpretation of the categorical imperative. For instance, one well-known interpretation is known as the "Humanity Formula", according to which human beings should never be treated as *merely* means to an end but always as ends-in-themselves (Johnson 2014; Kant, Kant 1787/1998). In line with the emerging evidence, we believe this notion that human beings are ends-in-themselves, and should never be used for instrumental purposes without regard for their humanity, is linked with the empathic and moral sentiments associated with the Phenomenal stance and its underlying neural signature (see Figure 1e).

According to utilitarian thinking, which is famously exemplified by the work of John Stuart Mill, the right moral action is that which maximizes the aggregate happiness (Mill, Mill 1861/1998). Utilitarian ethics has generally been associated with explicit and deliberate forms of instrumental reasoning. These can either be social in nature, such as interpersonal manipulative skills, or analytic in nature, such as regarding humans as statistical objects in order to calculate benefits and risks. There is thus a utilitarian tendency to instrumentally manipulate humans which is not only absent from, but universally discouraged by, deontological thinking, especially when one understands the categorical imperative in terms of the Humanity Formula.

Coupling these philosophical tenets with the empirical work reviewed above gives rise to the following conjectures: First, utilitarian thinking in these footbridge type moral dilemmas should be facilitated by a natural inclination to adopt both the Intentional stance and Physical stances, with the latter having a stronger effect. Second, and in direct contrast, the tendency to support deontological judgments in these same moral dilemmas should be associated with an inclination to step into the Phenomenal stance – which facilitates a tendency to connect with and respect the humanity in others. If these hypotheses are correct, then only a particular type of emotional thinking should relate to deontological responses – namely concern relating to the moral patiency of others.

The following falsifiable empirical predictions follow from these observations: First, measures of emotionality in general, such as personal distress and fear, should be unrelated to deontological responses in these scenarios. This is because these emotions are automatic and emerge from primitive limbic areas of the brain, whereas prosocial

and moral sentiments linked with Phenomenal stance are controlled and emerge from the DMN (e.g., Jack et al. 2012; Lindquist et al. 2012; Powers et al. 2015; Rameson et al. 2012). It is these latter sentiments in particular, as opposed to more primitive emotions, which should encourage empathic forms of connection with others. Second, emotionally disengaged measures of instrumental social reasoning (e.g., Theory of Mind) should either negatively predict deontological responses to these dilemmas or bear no relationship. This is because these are skills which do not require connecting with the humanity in others; indeed, in some instances they facilitate the manipulation of others (Blair et al. 1996). Third, measures of analytic reasoning should negatively predict deontological responses in these sorts of dilemmas (i.e., positively predict utilitarian responses).

There is emerging evidence in favor of these hypotheses. Many studies have shown that utilitarian thinking in these dilemmas is predicted by deficits in prosocial sentiments, rather than measures of emotionality or those associated with the Intentional stance. Instead, utilitarian decisions in these dilemmas are associated with psychopathic personality traits (Bartels and Pizarro 2011; Jack et al. 2014; Koenigs et al. 2012; Patil and Silani 2014). This is intriguing because psychopaths show both a reduced capacity to step into the Phenomenal stance (reduced empathic concern) and an increased capacity to step into the Intentional stance (Blair et al. 1996). Causal evidence for this view is provided by work demonstrating that experimentally increasing feelings of empathy leads to increased deontological decision making (Conway and Gawronski 2013).

There is also support for the hypothesis that individual differences and experimental inductions of various forms of controlled reasoning (e.g., mathematical, analytic; the physical stance) increase utilitarian responses (Paxton et al. 2012) to these dilemmas (but see Jack et al. 2014) and increase antisocial behavior in general (Small et al. 2007; Wang et al. 2014; Zhong 2011). Taken together, this work supports the view that treating humans as merely means to an end is facilitated by an attenuated ability to step into the Phenomenal stance and/or a natural inclination to step into the Intentional and Physical stances.

But why should such abstract moral rules as "never treat humans as means to an end, but always ends-in-themselves" be associated with the empathic sentiments instantiated by the DMN? One tentative interpretation is that higher levels of empathy and compassion incline us to connect with the humanity in others; that is, to respect the dignity that is intrinsic to all human beings (Jack et al. 2014). In doing so, we may be less likely to treat other human beings as instruments, mere means to an end.

In line with this, several different studies have shown decreased activation throughout the DMN while participants viewed pictures of individuals who were of instrumental value or viewed as less than fully human (Harris and Fiske 2006, 2007; Harris et al. 2014). Other researchers have shown that damage to key areas of the DMN involved with abstract emotional representation and emotional regulation (Roy et al. 2012) increases utilitarian responses to these dilemmas (Koenigs et al. 2012). Moreover, damage to brain areas that 'switch' between the DMN and TPN (Menon and Uddin 2010; Sridharan et al. 2008) results in a failure to fully engage the DMN during moral decision making, which corresponds with increased utilitarian responses to these footbridge dilemmas (Chiong et al. 2013). These findings are in line with the view that failure to fully engage the DMN facilitates the tendency to treat others as instruments.

We are not claiming that these results suggest that normatively desirable moral cognition requires that we spend all of our time in the Phenomenal stance while entirely ignoring either the Intentional and Physical stances. Mature and flexible ethical thinking most likely emerges from the ability to efficiently oscillate between these stances and their underlying neural networks. Moreover, the evidence reviewed here is particular to a certain class of moral dilemmas involving certain sorts of agents, namely footbridge type dilemmas with humans. Future work is wanted to determine how these personality traits relate to other dilemmas with other types of agents.

Lastly, as we emphasized above, we are not here to defend Kant and deontological ethics against Mill and utilitarian ethics. The details of Kant's ethics are unimportant for our argument. We appeal to Kant because his ethical system, and in particular the "Humanity Formula" of his categorical imperative, seems to be motivated by an appreciation for the sanctity of human life. In light of the reviewed philosophical theory and empirical evidence, we believe this appreciation is most closely linked with the phenomenal stance. This is something that individuals who endorse utilitarian judgments in these scenarios seem to (intentionally or inadvertently) fail to appreciate. In light of evidence that utilitarian responses to these and other sets of moral judgments do not reflect an "impartial concern for the greater good" (Kahane et al. 2015), it seems that treating humans as means to an end is not primarily driven by moral concern to increase the aggregate happiness of those who would experience the aggregated increase.[18]

4 X-Phi Reconsidered

In this section, we highlight differences between our approach to X-Phi and that of most other X-Phi projects. This is important because although other projects have argued for an inclusive relationship between cognitive science and philosophy, the philosophical import of such a synthesis is still a matter of debate. We contend that these concerns are not applicable to our approach. We further discuss how empirically

[18] Both this study and Kahane (2015) raise a larger question, which is whether the 'utilitarian responses' to these dilemmas reflect genuine utilitarian reasoning, e.g. since such reasoning is supposed to be driven by "impartial concern for the greater good". Clearly the tendency to make such responses does not capture all elements of utilitarian reasoning. However, it seems there can be little argument that utilitarian ethics would push more strongly than deontological ethics towards the responses in question. Similar concerns may be raised about the degree to which 'deontological responses' to footbridge dilemmas properly capture deontological reasoning (see study 2 of Jack et al. 2014). We have no argument with the view that a fully articulated utilitarian approach to ethics will incorporate elements from all the stances. Indeed, it must be so, since under our framework 'impartial concern', being a moral sentiment, must originate from the phenomenal stance and cannot be a product of the other stances. The key point is that both Utilitarianism and the so-called 'utilitarian responses' to footbridge dilemmas are similar in the respect that both place a much stronger weight on a detached and objective calculus than 'deontological' approaches or responses. Further, the claim that utilitarian perspectives rely more strongly (than deontological approaches) on analytic reason (or in Kant's terms 'theoretical reason' i.e. the Physical stance) is so obvious it is effectively beyond question; and it is similarly clear that Kant was explicitly determined to resist this heavy reliance on theoretical reason when he formulated his approach to ethics. Empirical measures (such as responses to trolley dilemmas) are inevitably artificial and limited proxies for full blown philosophical approaches, however such limitations do not necessarily render the findings invalid or uninformative provided the interpretation does not rely on a simple identification between approaches and responses, and takes heed of such limitations.

falsifiable hypotheses pertaining to the relationship between our psychology and opposing philosophical positions can immediately and directly enrich our understanding of the philosophical issues in ways that most other projects, including traditional forms of philosophical analysis, cannot. Thus, we claim that cognitive science offers a unique and valuable way to advance philosophical understanding, and hence is indispensable for moving the field of philosophy forward, at the same time as advancing cognitive science.

Many researchers have already demonstrated the benefits of using empirical methods to examine philosophical issues. For instance, we now have a considerably enriched understanding of the psychology of intuitions (Nado 2014). Notably, much of this work shows how intuitions are often sensitive to factors that are not normative (i.e. truth-tracking) and hence are best described as cognitive biases (Andow 2015; Byrd 2014; Livengood et al. 2010; Pinillos et al. 2011; Schwitzgebel and Cushman 2015). This work also indicates that philosophers tend to more carefully 'check' their intuitions with deliberate, reflective thought than do non-philosophers – a finding which is hardly surprising given the goals of philosophical education, but nonetheless significant and gratifying to see empirically supported (e.g., Andow 2015; Byrd 2014; Livengood et al. 2010). Some influential work in X-phi has also shed new light on the psychology of social and ethical reasoning, in particular judgements of intentionality and responsibility (Knobe 2003; Shaun and Knobe 2007). These findings, some of which are quite counter-intuitive, clearly make a significant contribution to cognitive science. We believe they are also informative to philosophy, at the very least because our understanding of such reasoning processes is clearly relevant to applied ethics. However, some have expressed skepticism about the philosophical relevance of these findings (e.g., Cappelen 2012; Devitt 2012; Kauppinen 2007; Sosa 2007). More specifically, it seems there is still trepidation concerning *how* such empirical and cognitive work can directly inform philosophical thinking.

One reason for such concern might be attributed to the observation that much work in X-Phi focuses on issues which, generally speaking, are not thought to be psychological in nature. For instance, metaphysical questions, such as those surrounding free will and dualism, are typically regarded as inquiries into a mind-independent reality. If philosophers are primarily interested in apprehending metaphysical truths in this mind-independent sense, it is certainly a good question how cognitive science can inform such philosphical inquiries. Edouard Machery seems to adopt this sort of view (see also Papineau 2011). In a recent book, Machery (2017) argues that while cognitive science can be directly informative to certain goals of conceptual analysis, research efforts into these sorts of metaphysical issues (e.g., dualism) "should often be given up, and our philosophical interests should be reoriented toward issues that do not turn on such facts " (p1). This is because the 'facts' that Machery has in mind are metaphysical or modal 'facts'. However, our focus here is different. It isn't on the metaphysical 'facts' in themselves, but rather on the origin of the tension *between* which facts 'should' prevail (à la Kant and his skeptical method).

We are not directly concerned with shedding light, for example, on the nature of consciousness. Rather our first concern is to shed light on the philosophical problem *of* consciousness – the conceptual scheme which gave us the idea there was something that we might be able to study. Put another way, we are not trying to answer the question (e.g. Does consciousness have a biological basis?), but rather to refine the

question (e.g. Does it make sense to ask if consciousness has a biological basis?).[19] We think 'facts', of a different sort, are of interest to understanding philosophical *problems* or *disputes* – namely 'facts' about our cognitive structure. If philosophical disputes – whether metaphysical or not – are hypothesized to (partially) emerge from our cognitive structure, then the philosophical utility of cognitive science becomes more clear. Indeed, the philosophical utility of cognitive science to better understanding the nature of philosophical disputes should not even appear suprising to anyone with a passing philosophical education, given that Kant influentially provided an essentially psychological theory of the origin of such disputes, and did so long before psychology emerged as a displine that could be distinguished from philosophy.

Scholl (2007) advances a similar view to ours, where he cogently argues that identifying the psychological mechanisms driving intuitions about object persistence and space can speak to the origin of related metaphysical claims (e.g., space and time). While we enthusiastically applaud Scholl's (2007) work, there are differences between his account and ours that are worth bringing to the fore. First, although Scholl (2007) attempts to ground metaphysical beliefs about space and time in our psychology, he does not draw parallels between this sort of empirical approach on the one hand, and traditional philosophical analyses attempting to ground these beliefs in our psychology on the other. Hence, Scholl (2007) regards cognitive science as being able to bridge the "disparate" (p.1) fields of philosophy and psychology, rather than constituting a form of philosophy itself. In contrast we argue that our approach should be understood as philosophy on the clear parallels between it and historically influential proto-psychological philosophical analyses (Section 3). Second, Scholl (2007) does not discuss how tension between psychological processes might give rise to competing metaphysical beliefs. This is central to our account. Third, Scholl (2007) suggests that work in cognitive science can give us reasons to relinquish certain intuitions/beliefs when they conflict with each other. In contrast, we suggest that better understanding the cognitive origins of conflicting philosophical views actually provides compelling reasons *not* to eliminate one view for another.[20] We highlight these differences merely to show how our approach differs from Scholl's (2007) work, not to undercut its philosophical significance.

Others have expressed hesitant optimism when it comes to the philosophical significance of cognitive science, whether it is used to study metaphysical issues or not. In a recent book chapter titled "Experimental philosophy is cognitive science,"

[19] Put briefly, if you mean consciousness in something like the sense Ned Block (2005) calls 'access consciousness', then yes, that can be turned into a structured and meaningful scientific project (Jack and Shallice 2001). If you mean "What is the biological basis of phenomenal consciousness?", then we contend that the question can't be answered. Not because metaphysical dualism is true, but because the question itself is incoherent. It is like asking "What is the semantic structure of an orange?" - a category mistake (see the conclusion of section 3.1). In other words, adopting the Physical stance to try and understand the 'what it is like' aspect of consciousness is not just starting off on the wrong foot, but it is signing up for the wrong race.

[20] The differences in how we and Scholl (2007) handle conflicting beliefs may be traced to the metaphysical issues each of us addresses. We are addressing issues which, we have argued, emerge from a tension between two fundamentally different modes of understanding the world, one which is analytic and another which is empathic, and both of which are essential for successfully navigating the world, albeit different aspects of that world. In contrast, intuitions about space, time and object persistence do not seem to emerge from such a fundamental tension, but might be traced to more nuanced tensions between different forms of analytic thinking (i.e., different types of representation associated with TPN activation and DMN suppression) which are both aimed at the same sort of understanding.

Knobe (2016) argued that we must construct a novel metaphilosphical framework in order to fully appreciate the philosphical import of such work. Using the metaphilosphical framework for traditional conceptual analysis is ill-suited, Knobe argues, because X-Phi does not try to provide the 'same sort of thing' as conceptual analsyis (e.g., a theory of concept *x*, or the necessary and sufficient conditions for determining concept *x*). We agree with Knobe that most work in X-Phi does not aim to provide the "same sort of thing" as traditional conceptual analysis. And we further agree that the philospohical import of such work can become more transparent by constructing a novel metaphilosphical framework.

However, to claim that a new metaphilosophical framework is required for appreciating the philosophical import of X-phi *because* "experimental philosophy is cognitive science" is, in some sense, to claim that it is exclusive; that is, to starkly distinguish X-phi (or cognitive science) from philosophy. We believe this sort of claim is premature because our type of work in X-phi *does* use cognitive science to provide exactly the 'same sort of thing' that the undeniably philosophical treatments discussed above aimed to provide – the illumination of a tension in us. Hence, our X-phi work is recognizably work in cognitive science, but not in an exclusive sense – it is also recognizably work in philosophy. Again, that is the thesis of this essay.

Because we regard the problems as emerging from a tension in us, rather than from an inherent inability to grasp facts about a mind-independent world, we regard ourselves as practicing the (Kantian) skeptical method. Consequently, we do not study individual differences in cognition in order to support one philosophical view over another. Instead, we appeal to philosophical judgments and individual differences *as data* that might illuminate the neural and psychological origins of philosophical issues, including their competing views. By mapping philosophically relevant cognitive stances onto opposing philosophical viewpoints, and using individual difference measures to assess one's tendency to adopt these stances, we are able to test empirically falsifiable hypotheses designed to reveal how our psychology inclines individuals to one or another philosophical view. And because we make hypotheses about the relationship these psychological processes share with each other, the results can speak to the tension between competing viewpoints.

The logic behind our own approach of emphasizing various individual difference measures can be more fully explicated as follows: We maintain that fundamentally opposed philosophical worldviews arise as a result of features of our shared cognitive structure. In other words, the existence of 'neurologically grounded antinomies' can be perceived (or felt) by all, or at least most, of us. Individual variation in certain cognitive capacities reflects the tendency for individuals, when forced to make a choice, to adopt one cognitive mode over another. For example, an individual who scores in the top quartile on the IPT (i.e., physical reasoning) and in the lowest quartile on the IRI-EC (i.e., empathic concern) is an individual who tends to adopt the physical stance more readily than the phenomenal stance, even when faced with stimuli that bias most people towards adoption of the phenomenal stance. Which stance the individual privileges will determine whether they are more likely to endorse one or another of the opposing worldviews we have discussed, when they are faced with questions that force such a decision. Hence, we are not in the business of trying to harmoniously reconcile the divergent views between the folk and the majority of philosophers. We believe there is sufficient evidence supporting the claim that we are cognitively constrained from ever

reconciling the opposing views, which are each supported both by the folk and by philosophers – albeit in different proportions.[21]

This is a major point of departure from most work in X-Phi, which presupposes that the divergent views evidenced between the folk and philosophers – and even among philosophers themselves – can be harmoniously reconciled with an ultimately unified and naturalistic worldview. These projects aim to explain 'away' metaphysically peculiar beliefs (e.g., intuitive dualism and intuitive libertarianism) as the product of epistemically defective psychological processes, such as emotions,[22] (Greene 2011; Nichols 2014), intuitive biases (Pennycook et al. 2012), cleverly designed intuition pumps (Dennett 1984, 1991, 2013),[23] or philosophical ignorance (Devitt 2012). These sorts of 'debunking' approaches suffer from at least two serious shortcomings.

First, they run the risk of inaccurately characterizing the psychological processes underlying opposing philosophical worldviews, and thus the philosophical worldviews themselves. Most of these approaches adopt a dual-process model of cognition (Evans and Stanovich 2013; Kahneman and Tversky 1972). On this view, beliefs and judgments emerge either immediately from quick and automatic processes in response to stimuli (Type 1 processes) or such judgments are subsequently 'checked and revised' by controlled, deliberate thought (Type 2 processes). The first sort of processing is often characterized as intuitive, primitive and error-prone, where the second is often characterized as rational, sophisticated and error-correcting. Emotions and empathy are often conflated with Type 1 processing, and analytic thinking with Type 2 processing. However, as we mentioned in section 2.1, this view is inconsistent with a growing body of evidence in neuroscience.

We have advanced an account according to which Type 1 and Type 2 processes exist within both the DMN and TPN, which means that certain emotional and analytic processes are automatic and intuitive on the one hand, whereas others are deliberate and controlled on the other (Friedman et al. 2015). Several neuroimaging studies have demonstrated that various types of empathic thinking are controlled Type 2 process (Morelli and Lieberman 2013; Rameson et al. 2012). We have provided evidence above (section 3.1, with Nagel) that empathic concern is associated with belief in dualism. Hence, researchers who characterize dualism (and similar beliefs which do not fit into a naturalistic worldview) as emerging from 'primitive' or 'error prone' thinking are inaccurately characterizing the cognitive processes associated with this belief, as well as the belief itself. This could inadvertently motivate us to view individuals who hold to such beliefs as primitive thinkers, which could have deleterious ethical effects on how such individuals are perceived (Haslam 2006).

Our approach regards such metaphysical beliefs as emerging from a tension between two sorts of reflective thinking, each of which is necessary for effective function. This is one motivation behind our suggestion that we learn to live with the conflict, rather

[21] It is not surprising that a greater proportion of philosophers endorse physicalism, whereas a greater percentage of the folk endorse dualism, given the findings cited earlier which indicate that philosophers tend to employ analytic reasoning more than the folk when considering their intuitions.

[22] "Metaphysicians rarely train their students to have heightened emotional sensitivities. "(Nichols 2014, p. 738)

[23] To be sure, we are not arguing that intuition pumps don't exist, or fail to influence, philosophical judgments. Rather, we want to highlight a critical difference between the exogenous influences of carefully crafted thought experiments, and the endogenous influence of our neural architecture on our intuition.

than attempting to dispose of it – because forfeiting beliefs which do not fit into a naturalistic worldview would seem to inevitably involve demeaning the cognitive processes that drive such beliefs. The more pressing reason is that our work pushes us to the hypothesis that effective ethical judgment is inevitably tied to the adoption of a dualistic world view.[24] In line with this hypothesis, we have previously noted that certain types of ethical judgment, in particular judgments associated with higher levels of empathic concern, are very hard to provide a rational basis for without appealing to dualistic notions such as talk about the 'soul' (Jack et al. 2014).

In further support of this, we have demonstrated that although spiritual and religious worldviews are negatively predicted by various sorts of analytic reasoning (Gervais and Norenzayan 2012; Pennycook et al. 2012; Shenhav et al. 2012) they are even more strongly positively predicted by various sorts of moral sentiments, especially empathic concern (Jack et al. 2016). It is worth noting that the positive effects of empathy on religious belief were strongest among the most analytically inclined participants, suggesting that individuals who have developed both sorts of reflective cognition (analytic and empathic) are more adept at discerning their unique affordances, an insight that can be traced to Kant. More specifically, we suggest that while highly analytic individuals recognize that certain religious/spiritual claims are empirically suspect (e.g., the world was created in 7 days), they may also recognize that certain supernatural beliefs, which P.F. Strawson would describe as 'panicky metaphysics', help cultivate social and moral insights by resonating with a form of understanding that is fundamentally opposed to naturalistic modes of thinking (Jack et al. 2016)

Second, these debunking approaches cannot adequately explain why such beliefs persist under the reflective scrutiny of expert philosophers. Recent research has shown that both philosophers and non-philosophers are susceptible to the same biases (Schwitzgebel and Cushman 2015), which suggests that dual-process theory does not adequately account for differences in philosophical worldviews. However, interpreting results through a theory of cognition that is not wholly accurate is not, by itself, a major cause of concern. The results can always be reinterpreted using another theory of cognition that more accurately reflects the way the brain is organized (as we have done with some of Joshua Greene's work, see section 2.2). What we regard as a major cause of concern is that many projects in X-Phi simply presuppose the truth of a metaphysical thesis and then use X-Phi to further support that view, or explain away conflicting views. Although this may seem like a caricature of (experimental) philosophy, at least two other prominent figures in philosophy have expressed similar views. David Hume claimed that "reason is, and ought only to be the slave of the passions" (Hume 1738/2012). And William James writes:

[24] The thought here is that the process of reliably generating normatively correct judgments requires us to adopt the phenomenal stance as part of our deliberations. This engages the 'morally reactive attitudes' and causes us to represent individuals as possessing free will and phenomenal consciousness. Note that we do not endorse the hypothesis that effective ethical judgment is tied to an idealistic world view, since we view effective ethical judgment as requiring a balancing of considerations generated by different stances (Jack et al. 2014; Rochford et al. 2016). Our pragmatic view is that while we should be dualists for the purposes of making ethical judgments, we should be monists when we are concerned with making judgments about the physical structure of the world.

> "[T]he history of philosophy is to a great extent that of a certain clash of temperaments. Undignified as such a treatment may seem to some of my colleagues, I shall have to take account of this clash and explain a good many of the divergencies [sic] of philosophers by it. Of whatever temperament a professional philosopher is, he tries when philosophizing to sink the fact of his temperament. Temperament is not conventionally recognized reason, so he urges impersonal reasons only for his conclusions. Yet his temperament really gives him a stronger bias than any of his more strictly objective premises. It loads the evidence for him one way or the other, making for a more sentimental or a more hard-hearted view of the universe…"
>
> (James 1906/1975, p. 2)

James went on to bifurcate temperament into the *tender-minded* and *tough-minded*. He describes the tender-minded as 'idealistic', 'religious', and 'free-willist'; the tough-minded as 'materialistic', 'irreligious', and 'fatalistic' (p. 4). These are opposed temperaments associated with opposed philosophical inclinations – and the similarity that his list has to the predictions emerging from the observation of the neural tension between the DMN (Phenomenal stance) and TPN (Physical stance) is notable. James even writes that the "[T]he tender feel the tough to be unrefined, callous**,** or brutal" (p. 5). This is quite a prescient insight in light of our results above. James identifies tender-mindedness with idealism, whereas the work outlined here identifies it with dualism; however, both these philosophical viewpoints share the feature of privileging (experiential) mind over the hegemony of materialism (e.g., the Phenomenal stance over the Physical stance).

We emphasize the above passage because James, like Kant, was acute to the tension there is between "facts and principles", where facts are aligned with the empirical/theoretical (TPN activation and DMN suppression; Physical stance) and principles are aligned with the moral/ethical (DMN activation, TPN suppression; Phenomenal stance). Even if one disagrees with our drawing a parallel between James' division of facts and principles with the anti-correlated empirical/theoretical cognitive processes instantiated by the TPN on the one hand, and the moral/ethical sentiments instantiated by the DMN on the other, it cannot be denied that James was tuned in to the tension there is between facts about our psychology (e.g., temperament) and our philosophical worldviews. Given James' status and influence on modern philosophy and psychology, it is rather odd that others have not paid attention to this insight and done more to incorporate individual difference measures into their experimental projects as means of systematically testing hypotheses about the origins of philosophical beliefs.

Our analysis of three different philosophical disputes as arising from one feature of our cognitive structure (i.e. the tension between the DMN and TPN) may raise some questions because these disputes differ in form in various ways. Do these differences undermine our claim that they each arise from the same underlying feature? We do not believe so. Undoubtedly many factors, including historical and cultural factors, influence the form of philosophical debates. Further, these debates are shaped by what is perceived to be at stake by the community. For instance, debates in ethics appear to have taken a fundamentally different form from the debates about consciousness and free will. In the case of utilitarian vs. deontological ethics what is at stake is which

theoretical framework can best inform practical ethical decision making, whereas in the cases of free will and the mind-body problem what is at stake is the fundamental nature of the mind. Hence, philosophers don't talk about utilitarian ethics as causing or constituting deontological ethics because it is recognized that these systems of ethics are human constructions. On the other hand, in discussions of free will and consciousness, we are discussing the nature of the phenomenon in itself, therefore it certainly makes sense to suggest that neural processes cause or constitute human experience and free will.

Nonetheless, our framework may be able to shed some further light on the distinct forms that debates about the mind-body problem and free will have taken. In particular, we hypothesize that there is a greater tension, and hence a deeper conceptual divide, between experience and neural processes on the one hand (Phenomenal vs. Physical stances), than there is between freely willed action and determinism (Phenomenal vs. Intentional stances).

For the problem of free will, since the conceptual divide is less pronounced, compatibilism has emerged as a common strategy for reconciling the competing viewpoints. In practice, this strategy has involved arguing that incompatabilism is untenable or fails to provide an adequate account even on its own terms, coupled with limited efforts explaining that intuitions of incompatibilism rest on some kind of misunderstanding.

For the mind-body problem, the tension between perspectives is greater. Historically, this encouraged philosophers to pick a side, such that they might adopt either idealism or physicalism. As science increasingly progressed, including making inroads on the mind, idealism became less popular and the physical nature of matter and mind became harder to resist. However, the perception that something was 'left out' of physicalist accounts of the mind nonetheless remained too salient to simply gloss over or push aside in the same manner as intuitions about free will. This has led to at least one of three strategies that are commonly employed to address this concern. The first consists of holding to a nuanced form of dualism (i.e. property dualism or dual aspect theory), often accompanied by claims about bridging principles or laws that can mediate between the two perspectives, which softens the essentially non-explanatory nature of such accounts. A second consists of providing more elaborate arguments to deny that any sense can be made of the intuition that an explanatory gap persists, for instance by undermining faith in introspection and arguing that some kind of intentionally fabricated – as opposed to neurologically grounded – cognitive illusion is occurring (e.g. 'intuition pumps'). The third strategy is to admit that an apparent explanatory gap remains, but to optimistically resist dualism in the hope that scientific advances will eventually close the gap by producing a theory of consciousness.

Hence, it would appear that the greater salience of the explanatory gap that is present for the problem of consciousness (in contrast to the problem of free will), which according to our account arises from the structure of cognition, explains the greater lengths that philosophers have gone to in their attempts to provide a satisfactory resolution in favor of one metaphysical position or another. However, to be clear, we believe our work indicates that such efforts are ultimately futile, and will ultimately be seen as sophistry, although well intentioned rather than malicious sophistry. Indeed, we would go even further than this, and question why analytic philosophers have expended so much effort in trying to argue for one metaphysical position over another. It strikes

us that a highly dubious assumption underlies such efforts, namely the view that the tools of analytic philosophy (e.g., conceptual analysis) are well suited for determining what is metaphysically true.[25] In our view, the scientific method, i.e. putting forward hypotheses about nature and then testing them empirically, represents the best method for determining what is metaphysically true. Conceptual analysis cannot provide any direct support for the truth or falsity of any particular metaphysical claim.[26] On the contrary, we would suggest that what conceptual analysis does directly inform is precisely the nature of human understanding. Concepts are, after all, merely the mental constructs that underlie and support human understanding. Our conceptual structure reveals the structure of cognition, regardless of whether all the concepts concerned demonstrate correspondence with the world.

Hence the exercise of claiming that the way we understand the mind implies that certain things are true about the mind (e.g. Fodor's 'Language of Thought' hypothesis, or the claim that phenomenal concepts succeed in referring to natural properties of mental states) is not, and should never have been taken as, evidence which speaks directly to the actual constitution of the mind. Rather, it is most directly evidence about the cognitive structures we have evolved to understand minds. Hence, it does not strike us as a reasonable view that science, the only true route to metaphysics, should be expected or encouraged to give undue weight to the Language of Thought hypothesis or the hypothesis there is a biological basis for phenomenal consciousness - any more than science should strive to demonstrate the existence of an immaterial soul. Such strivings might be likened to the view that physicists should expend additional efforts attempting to prove true the false beliefs of naive or folk physics. In summary, while our primary focus in this paper has been to argue that some projects in cognitive science also constitute works of philosophy, we believe the converse to also be true. The philosophical method of the analysis of conceptual tension may be seen as an excellent tool for pinpointing tensions in human understanding, which can be readily transformed into hypotheses in cognitive science. Hence, some work in philosophy also makes an invaluable contribution to the discipline of cognitive science.

[25] Machery (2017) has recently claimed that philosophers are 'modally immodest' by supposing that analyzing certain situations or intuitions can reveal necessary metaphysical truths. We agree. However, Machery did not arrive at this conclusion by appealing to any sorts of facts about our cognitive structure. Instead, he essentially argues that because we could never possibly ascertain the existence of necessary metaphysical truths, (most) philosophers have to change the way they philosophize, or at least appreciate the (modal) limitations of certain philosophical practices. Although there are other differences between our accounts, we reference Machery merely to illustrate that other X-Phi proponents see how problematic is to move from consulting one's intuition's, or reflective analyses about thought experiments, to metaphysical claims about the world (and other possible worlds).

[26] Conceptual analysis is of course well suited to examining the underlying assumptions of successful scientific frameworks and/or theories, and may borrow from the epistemic credibility of those frameworks in order to support claims that nature is structured in a particular way. However, we cannot see how conceptual analysis can be taken as having any epistemic credibility except by this indirect route, and hence the credibility of any metaphysical claims put forward on the basis of conceptual analysis ultimately derive entirely from empirical testing.

5 Conclusion

We have argued that our experimental project shares a close affinity with classical forms of analysis in philosophy, the major difference being that our epistemic basis relies on data from the cognitive sciences, not on introspection or logical analysis. This evidence complements these philosophically traditional analyses, and points to a similar conclusion, but does so in ways that introspection and logical analysis alone could not possibly achieve. The type of philosophical analysis afforded by our approach seeks to provide a descriptive account of the tension *in us* that gives rise to incommensurable philosophical beliefs. We traced this tension to our neurology and the antagonistic relationship between two anatomically distinct and functionally inhibitory neural networks, one underlying empathic attachment and moral sentiments (Default Mode Network; DMN), and another underlying empirical reasoning and analytic thought (Task Positive Network; TPN).

We want to close by emphasizing that the prospect of grounding philosophical problems in our neurology has the potential to mitigate a certain futility of purpose that can often be seen when we take a more distanced view of the back and forth of philosophical discourse. By illuminating the cognitive forces behind certain problems, we can learn to appreciate that the inability to reconcile competing worldviews is due to endogenous cognitive constraints rather than philosophical ignorance or wishful thinking. We may also learn that some of these cognitive forces are laudable and not worth stifling (e.g., Baumeister et al. 2009; Vohs and Schooler 2008). Having recognized this, we might rethink the motivation to banish metaphysical dualism in favor of reductive physicalism, a worldview that shares a relationship with psychopathic tendencies. It is not our claim that all philosophical problems can be grounded within our theoretical framework, or that our apparent inability to simultaneously deploy the neural networks underlying moral sentiments and physical reasoning (broadly construed) is the answer to every intractable philosophical question. We suspect that other competing neurological and psychological processes can illuminate the intractable nature of some other philosophical problems and paradoxes. We look forward to work adopting this schematic approach. Nevertheless, we believe that a fuller understanding of the world, and our place in it, can be cultivated by respecting the exclusivity of each of these incommensurable domains of understanding and using them in the appropriate context.

Finally, we want to highlight two different ways in which beliefs and attitudes may be judged (ir)rational and how these relate to our project (c.f., Nozick 1981). On one view, a belief, choice or attitude is rational because it increases utility or happiness or good fortune. Another sense in which a belief or attitude can be rational depends on how well it 'fits with' or 'hangs together' with other beliefs and attitudes. The better it fits, the more rational it becomes to incorporate. In the first sense, it appears to be rational, in some contexts, to hold to every one of the positions that we have discussed (e.g. physicalism and dualism, compatibilism and incompatibilism, utilitarianism and deontology). However, it is an emphasis on the latter sense that encourages many philosophers to mistakenly believe that only one of two incommensurable viewpoints *must* be correct – that is, holding both of them seems irrational in this sense. We have gone to some length to explain why this view is, in these cases, mistaken. Our cognitive architecture predisposes us to form certain beliefs and attitudes that are contradictory, and thus irrational in this later sense, because they emerge from distinct and profoundly

opposed cognitive processes. Rather than give in to the inclination, which many philosophers evince, to dismiss or deride this type of irrationality, we believe philosophy would do better to celebrate it. Not only is it our uniquely human endowment, but also, it keeps philosophy in business.

Acknowledgements We would like to thank Chris Haufe, Joshua Knobe, Eddy Nahmias, Phillip Robbins and members of the MindsOnline Philosophy conference (2015) for helpful comments on prior versions of this manuscript. We would also like to thank Chin-Tai Kim for his many helpful discussions with J.P.F regarding Immanuel Kant's philosophy. Finally, we would like to thank Brent Strickland, reviewer [revealed reviewer] and one anonymous reviewer for insightful comments throughout the revision process.

References

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means visual imagery and moral judgment. *Psychological Science*, 23(8):861–868.

Amodio, D.M., and C.D. Frith. 2006. Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews. Neuroscience* 7 (4): 268–277. doi:10.1038/nrn1884.

Andow, J. 2015. How distinctive is philosophers' intuition talk? *Metaphilosophy* 46 (4–5): 515–538.

Andrew, J., M. Cooke, and S. Muncer. 2008. The relationship between empathy and Machiavellianism: An alternative to empathizing–systemizing theory. *Personality and Individual Differences* 44 (5): 1203–1211.

Andrews-Hanna, J.R. 2011. The Brain's Default Network and Its Adaptive Role in Internal Mentation. *The Neuroscientist*. doi:10.1177/1073858411403316.

Anticevic, A., M.W. Cole, J.D. Murray, P.R. Corlett, X.J. Wang, and J.H. Krystal. 2012. The role of default network deactivation in cognition and disease. *Trends in Cognitive Sciences* 16 (12): 584–592. doi:10.1016/j.tics.2012.10.008.

Auyeung, B., S. Wheelwright, C. Allison, M. Atkinson, N. Samarawickrema, and S. Baron-Cohen. 2009. The children's empathy quotient and systemizing quotient: Sex differences in typical development and in autism spectrum conditions. *Journal of Autism and Developmental Disorders* 39 (11): 1509–1521.

Babiak, P., and R.D. Hare. 2006. *Snakes in suits: When psychopaths go to work*. New York: Regan Books.

Bagozzi, R.P., W.J.M.I. Verbeke, R.C. Dietvorst, F.D. Belschak, W.E. van den Berg, and W.J.R. Rietdijk. 2013. Theory of Mind and Empathic Explanations of Machiavellianism: A Neuroscience Perspective. *Journal of Management*. doi:10.1177/0149206312471393.

Bandura, A. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3 (3): 193–209.

Baron-Cohen, S. 2002. The extreme male brain theory of autism. *Trends in Cognitive Sciences* 6 (6): 248–254.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. 2001a. The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry* 42 (2): 241–251.

Baron-Cohen, S., S. Wheelwright, A. Spong, V. Scahill, and J. Lawson. 2001b. Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of Developmental and Learning Disorders* 5 (1): 47–78.

Bartels, D.M., and D.A. Pizarro. 2011. The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121 (1): 154–161. doi:10.1016/j.cognition.2011.05.010.

Baumeister, R.F., E. Masicampo, and C.N. DeWall. 2009. Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin* 35 (2): 260–268.

Baumgartner, T., U. Fischbacher, A. Feierabend, K. Lutz, and E. Fehr. 2009. The neural circuitry of a broken promise. *Neuron* 64 (5): 756–770.

Beaty, R.E., M. Benedek, R.W. Wilkins, E. Jauk, A. Fink, P.J. Silvia, et al. 2014. Creativity and the default network: a functional connectivity analysis of the creative brain at rest. *Neuropsychologia* 64: 92–98.

Bellugi, U., L. Lichtenberger, W. Jones, Z. Lai, and M.S. George. 2000. I. The neurocognitive profile of Williams Syndrome: a complex pattern of strengths and weaknesses. *Journal of Cognitive Neuroscience* 12 (Supplement 1): 7–29.

Blair, R.J.R. 2005. Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition* 14 (4): 698–718.

Blair, J., C. Sellars, I. Strickland, F. Clark, A. Williams, M. Smith, and L. Jones. 1996. Theory of mind in the psychopath. *Journal of Forensic Psychiatry* 7 (1): 15–25.

Block, N. 2005. Two neural correlates of consciousness. *Trends in Cognitive Sciences* 9 (2): 46–52. doi:10.1016/j.tics.2004.12.006.

Bloom, P. 2004. *Descartes' baby: how the science of child development explains what makes us human*. New York: Basic Books.

Bressler, S.L., and V. Menon. 2010. Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences* 14 (6): 277–290. doi:10.1016/j.tics.2010.04.004.

Broyd, S.J., C. Demanuele, S. Debener, S.K. Helps, C.J. James, and E.J. Sonuga-Barke. 2009. Default-mode brain dysfunction in mental disorders: a systematic review. *Neuroscience and Biobehavioral Reviews* 33 (3): 279–296. doi:10.1016/j.neubiorev.2008.09.002.

Buckner, R., and D. Carroll. 2007. Self-projection and the brain. *Trends in Cognitive Sciences* 11 (2): 49–57. doi:10.1016/j.tics.2006.11.004.

Buckner, R.L., J.R. Andrews-Hanna, and D.L. Schacter. 2008. The Brain's Default Network: Anatomy, Function, and Relevance to Disease. *Annals of the New York Academy of Sciences* 1124 (1): 1–38. doi:10.1196/annals.1440.011.

Byrd, N. (2014). Intuitive and Reflective Responses in Philosophy. University of Colorado.

Bzdok, D., L. Schilbach, K. Vogeley, K. Schneider, A.R. Laird, R. Langner, and S.B. Eickhoff. 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function* 217 (4): 783–796. doi:10.1007/s00429-012-0380-y.

Cappelen, H. (2012). *Philosophy without intuitions*: Oxford University Press.

Carnap, R. 1955. Meaning and synonymy in natural languages. *Philosophical Studies* 6 (3): 33–47.

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.): The University of Chicago Press.

Caruso, E.M., and F. Gino. 2011. Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. *Cognition* 118 (2): 280–285.

Chalmers, D.J. 1996. Facing up to the problem of consciousness. *Toward a Science of Consciousness*: 5–28.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory*: Oxford Paperbacks.

Chiong, W., Wilson, S. M., D'Esposito, M., Kayser, A. S., Grossman, S. N., Poorzand, P., . . . Rankin, K. P. (2013). The salience network causally influences default mode network activity during moral reasoning. *Brain*, awt066.

Christ, S.E., D.C. Van Essen, J.M. Watson, L.E. Brubaker, and K.B. McDermott. 2009. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex* 19 (7): 1557–1566. doi:10.1093/cercor/bhn189.

Churchland, P.M. 1981. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*: 67–90.

Clark, C.J., J.B. Luguri, P.H. Ditto, J. Knobe, A.F. Shariff, and R.F. Baumeister. 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106 (4): 501.

Conway, P., and B. Gawronski. 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology* 104 (2): 216.

Corbetta, M., E. Akbudak, T.E. Conturo, A.Z. Snyder, J.M. Ollinger, H.A. Drury, et al. 1998. A common network of functional areas for attention and eye movements. *Neuron* 21 (4): 761–773.

Davis, M.H. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44 (1): 113.

Dehaene, S., and L. Cohen. 2007. Cultural recycling of cortical maps. *Neuron* 56 (2): 384–398. doi:10.1016/j.neuron.2007.10.004.

Dennett. (1984). *Elbow room: The varieties of free will worth wanting*: MIT Press.

Dennett. (1989). *The intentional stance*: MIT press.

Dennett. 1991. *Consciousness explained*. 1st ed. Boston: Little, Brown and Co..

Dennett. (2013). *Intuition pumps and other tools for thinking*: WW Norton & Company.

Denny, B.T., H. Kober, T.D. Wager, and K.N. Ochsner. 2012. A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience* 24 (8): 1742–1752. doi:10.1162/jocn_a_00233.

Descartes, R. (1641/1996). *Discourse on the Method: And, Meditations on First Philosophy*. Yale University Press.

Devitt, M. 2012. The role of intuitions in the philosophy of language. *Russell and Fara* 2012: 554–565.

Duncan, J., and A.M. Owen. 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences* 23 (10): 475–483.

Dutton, K. 2012. *The Wisdom of Psychopaths*. New York: Farrar, Straus and Giroux.

Dziobek, I., K. Rogers, S. Fleck, M. Bahnemann, H.R. Heekeren, O.T. Wolf, and A. Convit. 2008. Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders* 38 (3): 464–473.

Evans, J.S.B., and K.E. Stanovich. 2013. Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science* 8 (3): 223–241.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, 201610344.

Fox, M.D., and M.E. Raichle. 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience* 8 (9): 700–711.

Fox, M.D., A.Z. Snyder, J.L. Vincent, M. Corbetta, D.C. Van Essen, and M.E. Raichle. 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America* 102 (27): 9673–9678. doi:10.1073/pnas.0504136102.

Friedman, J., Jack, A. I., Rochford, K., & Boyatzis, R. (2015). Antagonistic Neural Networks Underlying Organizational Behavior. *Organizational Neuroscience (Monographs in Leadership and Management, Volume 7) Emerald Group Publishing Limited*, 7:115–141.

Gervais, W.M., and A. Norenzayan. 2012. Analytic thinking promotes religious disbelief. *Science* 336 (6080): 493–496. doi:10.1126/science.1215647.

Gleichgerrcht, E., T. Torralva, A. Rattazzi, V. Marenco, M. Roca, and F. Manes. 2013. Selective impairment of cognitive empathy for moral judgment in adults with high functioning autism. *Social Cognitive and Affective Neuroscience* 8 (7): 780–788.

Goel, V. 2007. Anatomy of deductive reasoning. *Trends in Cognitive Sciences* 11 (10): 435–441. doi:10.1016/j.tics.2007.09.003.

Greene, J. D. (2007). The secret joke of Kant's soul. *Moral Psychology: Historical and Contemporary Readings*, 359–372.

Greene, J.D. 2011. Social Neuroscience and the Soul's Last Stand. In *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*, ed. A. Todorov, S.T. Fiske, and D. Prentice. New York: Oxford University Press.

Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293 (5537): 2105–2108. doi:10.1126/science.1062872.

Greene, J.D., L.E. Nystrom, A.D. Engell, J.M. Darley, and J.D. Cohen. 2004. The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron* 44 (2): 389–400. doi:10.1016/j.neuron.2004.09.027.

Gunia, B.C., L. Wang, L. Huang, J. Wang, and J.K. Murnighan. 2012. Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management Journal* 55 (1): 13–33.

Haldane, J. 1989. Brentano's problem. *Grazer Philosophische Studien* 35: 1–32.

Harris, L.T., and S.T. Fiske. 2006. Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. *Psychological Science* 17 (10): 847–853. doi:10.1111/j.1467-9280.2006.01793.x.

Harris, L.T., and S.T. Fiske. 2007. Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience* 2 (1): 45–51.

Harris, T., V.K. Lee, B.H. Capestany, and A.O. Cohen. 2014. Assigning economic value to people results in dehumanization brain response. *Journal of Neuroscience, Psychology, and Economics* 7 (3): 151.

Haslam, N. 2006. Dehumanization: an integrative review. *Personality and Social Psychology Review* 10 (3): 252–264. doi:10.1207/s15327957pspr1003_4.

Hume, D. (1738/2012). *A treatise of human nature*: Courier Corporation.

Huth, A.G., W.A. de Heer, T.L. Griffiths, F.E. Theunissen, and J.L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532 (7600): 453–458.

Hyatt, C.J., V.D. Calhoun, G.D. Pearlson, and M. Assaf. 2015. Specific default mode subnetworks support mentalizing as revealed through opposing network recruitment by social and semantic FMRI tasks. *Human Brain Mapping* 36 (8): 3047–3063.

Iacoboni, M., M.D. Lieberman, B.J. Knowlton, I. Molnar-Szakacs, M. Moritz, C.J. Throop, and A.P. Fiske. 2004. Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage* 21 (3): 1167–1173. doi:10.1016/j.neuroimage.2003.11.013.

Jack, A. (2014). A scientific case for conceptual dualism: the problem of consciousness and the opposing domains hypothesis. In J. Knobe & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy* (Vol. 1).

Jack, A.I., and P. Robbins. 2012. The Phenomenal Stance Revisited. *Review of Philosophy and Psychology* 3 (3): 383–403.

Jack, A.I., and T. Shallice. 2001. Introspective physicalism as an approach to the science of consciousness. *Cognition* 79 (1–2): 161–196.

Jack, A.I., C.M. Sylvester, and M. Corbetta. 2006. Losing our brainless minds: how neuroimaging informs cognition. *Cortex* 42 (3): 418–421 discussion 422-417.

Jack, A.I., A.J. Dawson, K.L. Begany, R.L. Leckie, K.P. Barry, A.H. Ciccia, and A.Z. Snyder. 2012. fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage* 66C: 385–401. doi:10.1016/j.neuroimage.2012.10.061.

Jack, A.I., A.J. Dawson, and M.E. Norr. 2013. Seeing human: distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage* 79: 313–328.

Jack, A. I., Robbins, P., Friedman, J. P., & Meyers, C. D. (2014). More than a feeling: counterintuitive effects of compassion on moral judgment. *Advances in Experimental Philosophy of Mind* (Vol. 125): Continuum.

Jack, A.I., J.P. Friedman, R.E. Boyatzis, and S.N. Taylor. 2016. Why do you believe in God? Relationships between religious belief, analytic thinking, mentalizing and moral concern. *PloS One* 11 (3): e0149989.

Jack, A. I., Friedman, J. P., Luguri, J. B., & Knobe, J. (under revision). Consciousness and callousness: Distinct moral sentiments drive distinct metaphysically odd beliefs about the mind.

James, W. 1904. DoesConsciousness' Exist? *The Journal of philosophy, psychology and scientific methods* 1 (18): 477–491.

James, W. (1906/1975). *Pragmatism* (Vol. 1): Harvard University Press.

Johnson, R. (2014). Kant's Moral Philosophy. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Kahane, G. 2015. Sidetracked by trolleys: why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience* 10 (5): 551–560.

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

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Kahneman, D., and A. Tversky. 1972. Subjective Probability - Judgment of Representativeness. *Cognitive Psychology* 3 (3): 430–454.

Kant, I. (1787/1998). Critique of pure reason. In P. Guyer (Ed.): Cambridge University Press.

Kant, I. (1902). Prolegomena to any future metaphysics that can qualify as a science. Open Court Publishing.

Kauppinen, A. 2007. The rise and fall of experimental philosophy. *Philosophical Explorations* 10 (2): 95–118.

Kim, J. 1984. Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy* 9 (1): 257–270.

Kim, J. (2007). *Physicalism, or something near enough*. Princeton University Press.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63 (279): 190–194.

Knobe, J. (Ed.) (2016). *Experimental philosophy is cognitive science* (Vol. Blackwell).

Koenigs, M., M. Kruepke, J. Zeier, and J.P. Newman. 2012. Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience* 7 (6): 708–714. doi:10.1093/scan/nsr048.

Krendl, A.C., C.N. Macrae, W.M. Kelley, J.A. Fugelsang, and T.F. Heatherton. 2006. The good, the bad, and the ugly: an fMRI investigation of the functional anatomic correlates of stigma. *Social Neuroscience* 1 (1): 5–15. doi:10.1080/17470910600670579.

Krendl, A.C., T.F. Heatherton, and E.A. Kensinger. 2009. Aging minds and twisting attitudes: an fMRI investigation of age differences in inhibiting prejudice. *Psychology and Aging* 24 (3): 530.

Leibniz, G. W., & Montgomery, G. R. (2005). *Discourse on Metaphysics and the Monadology*: Courier Corporation.

Levine, J. (2000). Conceivability, identity, and the explanatory gap. Toward a Science of Consciousness Iii, 3-12.

Lieberman, M. D. (2007). The X-and C-systems. Social neuroscience: Integrating Biological and psychological explanations of social behavior, 290–315.

Lindquist, K.A., T.D. Wager, H. Kober, E. Bliss-Moreau, and L.F. Barrett. 2012. The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences* 35 (3): 121–143. doi:10.1017/S0140525X11000446.

Livengood, J., J. Sytsma, A. Feltz, R. Scheines, and E. Machery. 2010. Philosophical temperament. *Philosophical Psychology* 23 (3): 313–330.

Lockwood, P.L., G. Bird, M. Bridge, and E. Viding. 2013. Dissecting empathy: high levels of psychopathic and autistic traits are characterized by difficulties in different social information processing domains. *Frontiers in Human Neuroscience* 7: 760.

Machery, E. (2013). PodBean podcast. *Edouard Machery & Tony Sobrado on Consciousness*.

Machery, E. 2017. *Philosophy within its proper bounds*. Oxord: Oxford University Press.

Martin, A., and J. Weisberg. 2003. Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology* 20 (3–6): 575–587. doi:10.1080/02643290342000005.

Mason, M.F., M.I. Norton, J.D. Van Horn, D.M. Wegner, S.T. Grafton, and C.N. Macrae. 2007. Wandering Minds: The Default Network and Stimulus-Independent Thought. *Science* 315 (5810): 393–395. doi:10.1126/science.1131295.

Mather, M., J.T. Cacioppo, and N. Kanwisher. 2013. How fMRI can inform cognitive theories. *Perspectives on Psychological Science* 8 (1): 108–113.

Menon, V., and L.Q. Uddin. 2010. Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function* 214 (5–6): 655–667.

Meyer, M.L., S.E. Taylor, and M.D. Lieberman. 2015. Social working memory and its distinctive link to social cognitive ability: an fMRI study. *Social Cognitive and Affective Neuroscience* 10 (10): 1338–1347.

Mill, J. S. (1861/1998). *Utilitarianism* (R. Crisp Ed.). Oxford: Oxford University Press.

Molenberghs, P., H. Johnson, J.D. Henry, and J.B. Mattingley. 2016. Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews* 65: 276–291.

Morelli, S.A., and M.D. Lieberman. 2013. The role of automaticity and attention in neural processes underlying empathy for happiness, sadness, and anxiety. *Frontiers in Human Neuroscience* 7: 160.

Nado, J. 2014. Why intuition? *Philosophy and Phenomenological Research* 89 (1): 15–41.

Nagel, T. 1974. What Is It Like to Be a Bat. *Philosophical Review* 83 (4): 435–450.

Nakano, T., M. Kato, Y. Morito, S. Itoi, and S. Kitazawa. 2013. Blink-related momentary activation of the default mode network while viewing videos. *Proceedings of the National Academy of Sciences* 110 (2): 702–706.

Nichols, S. 2014. Process Debunking and Ethics*. *Ethics* 124 (4): 727–749.

Nowicki, S., & Duke, M. P. (2001). Nonverbal receptivity: The Diagnostic Analysis of Nonverbal Accuracy (DANVA).

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Papineau, D. 2011. What is x-phi good for? *The Philosophers' Magazine* 52: 83–88.

Patil, I., and G. Silani. 2014. Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology* 5: 501.

Paulhus, D., C. Neumann, and R. Hare. 2009. *Manual for the self-report psychopathy scale*. Toronto: Multi-health systems.

Paxton, J.M., L. Ungar, and J.D. Greene. 2012. Reflection and reasoning in moral judgment. *Cognitive Science* 36 (1): 163–177.

Pennycook, G., J.A. Cheyne, P. Seli, D.J. Koehler, and J.A. Fugelsang. 2012. Analytic cognitive style predicts religious and paranormal belief. *Cognition* 123 (3): 335–346.

Pinillos, N.Á., N. Smith, G.S. Nair, P. Marchetto, and C. Mun. 2011. Philosophy's new challenge: Experiments and intentional action. *Mind & Language* 26 (1): 115–139.

Powers, K. E., Chavez, R. S., & Heatherton, T. F. (2015). Individual differences in response of dorsomedial prefrontal cortex predict daily social behavior. *Social Cognitive and Affective Neuroscience*, nsv096.

Prabhakaran, V., J.A. Smith, J.E. Desmond, G.H. Glover, and J.D. Gabrieli. 1997. Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive Psychology* 33 (1): 43–63.

Price, C.J., and K.J. Friston. 1997. Cognitive conjunction: a new approach to brain activation experiments. *NeuroImage* 5 (4 Pt 1): 261–270. doi:10.1006/nimg.1997.0269.

Raichle, M.E., A.M. MacLeod, A.Z. Snyder, W.J. Powers, D.A. Gusnard, and G.L. Shulman. 2001. A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America* 98 (2): 676–682. doi:10.1073/pnas.98.2.676 98/2/676.

Rameson, L.T., S.A. Morelli, and M.D. Lieberman. 2012. The neural correlates of empathy: experience, automaticity, and prosocial behavior. *Journal of Cognitive Neuroscience* 24 (1): 235–245.

Rand, D.G., J.D. Greene, and M.A. Nowak. 2012. Spontaneous giving and calculated greed. *Nature* 489 (7416): 427–430.

Reniers, R.L., R. Corcoran, B.A. Völlm, A. Mashru, R. Howard, and P.F. Liddle. 2012. Moral decision-making, ToM, empathy and the default mode network. *Biological Psychology* 90 (3): 202–210.

Robbins, P., and A.I. Jack. 2006. The phenomenal stance. *Philosophical Studies* 127 (1): 59–85.

Rochford, K.C., A.I. Jack, R.E. Boyatzis, and S.E. French. 2016. Ethical leadership as a balance between opposing neural networks. *Journal of Business Ethics*: 1–16.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.

Roy, M., D. Shohamy, and T.D. Wager. 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences* 16 (3): 147–156. doi:10.1016/j.tics.2012.01.005.

Ryle, G. (1984). The concept of mind (1949). London: Hutchinson.

Sarkissian, H., A. Chatterjee, F. De Brigard, J. Knobe, S. Nichols, and S. Sirker. 2010. Is belief in free will a cultural universal? *Mind & Language* 25 (3): 346–358.

Schilbach, L., S. Eickhoff, A. Rotarskajagiela, G. Fink, and K. Vogeley. 2008. Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition* 17 (2): 457–467. doi:10.1016/j.concog.2008.03.013.

Scholl, B.J. 2007. Object persistence in philosophy and psychology. *Mind & Language* 22 (5): 563–591.

Schurz, M., J. Radua, M. Aichhorn, F. Richlan, and J. Perner. 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews* 42: 9–34.

Schwitzgebel, E., and F. Cushman. 2015. Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition* 141: 127–137.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Shariff, A.F., J.D. Greene, J.C. Karremans, J.B. Luguri, C.J. Clark, J.W. Schooler, et al. 2014. Free Will and Punishment A Mechanistic View of Human Nature Reduces Retribution. *Psychological Science* 25 (8): 1563–1570.

Shaun, N., and J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 41 (4): 663–685.

Shenhav, A., D.G. Rand, and J.D. Greene. 2012. Divine intuition: cognitive style influences belief in God. *Journal of Experimental Psychology: General* 141 (3): 423.

Shepherd, J., and J. Justus. 2015. X-Phi and Carnapian explication. *Erkenntnis* 80 (2): 381–402.

Shulman, G.L., M. Corbetta, R.L. Buckner, J.A. Fiez, F.M. Miezin, M.E. Raichle, and S.E. Petersen. 1997a. Common blood flow changes across visual tasks: I. Increases in subcortical structures and cerebellum but not in nonvisual cortex. *Journal of Cognitive Neuroscience* 9 (5): 624–647.

Shulman, G.L., J.A. Fiez, M. Corbetta, R.L. Buckner, F.M. Miezin, M.E. Raichle, and S.E. Petersen. 1997b. Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience* 9 (5): 648–663.

Small, D.A., G. Loewenstein, and P. Slovic. 2007. Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes* 102 (2): 143–153. doi:10.1016/j.obhdp.2006.01.005.

Sosa, E. 2007. Experimental philosophy and philosophical intuition. *Philosophical Studies* 132 (1): 99–107.

Sporns, O. 2014. Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience* 17 (5): 652–660.

Spreng, R.N. 2012. The fallacy of a "task-negative" network. *Frontiers in Psychology* 3: 145.

Spunt, R. P., & Lieberman, M. D. (2014). Automaticity, control, and the social brain. Dual-process theories of the social mind, 279–296.

Spunt, R.P., M.L. Meyer, and M.D. Lieberman. 2015. The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience* 27 (6): 1116–1124.

Sridharan, D., D.J. Levitin, and V. Menon. 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences* 105 (34): 12569–12574.

Strawson, P.F. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48: 1–25.

Sytsma, J., and E. Machery. 2009. Two conceptions of subjective experience. *Philosophical Studies* 151 (2): 299–327. doi:10.1007/s11098-009-9439-x.

Toplak, M.E., R.F. West, and K.E. Stanovich. 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition* 39 (7): 1275–1289.

Tversky, A., and D. Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5 (2): 207–232.

Van Overwalle, F. 2009. Social cognition and the brain: a meta-analysis. *Human Brain Mapping* 30 (3): 829–858.

Van Overwalle, F. 2011. A dissociation between social mentalizing and general reasoning. *NeuroImage* 54 (2): 1589–1599.

Van Overwalle, F., and M. Vandekerckhove. 2013. Implicit and explicit social mentalizing: dual processes driven by a shared neural network. *Frontiers in Human Neuroscience* 7: 560.

Vincent, J.L., G.H. Patel, M.D. Fox, A.Z. Snyder, J.T. Baker, D.C. Van Essen, et al. 2007. Intrinsic functional architecture in the anaesthetized monkey brain. *Nature* 447 (7140): 83–86. doi:10.1038/nature05758.

Vohs, K.D., and J.W. Schooler. 2008. The value of believing in free will encouraging a belief in determinism increases cheating. *Psychological Science* 19 (1): 49–54.

Wang, L., C.-B. Zhong, and J.K. Murnighan. 2014. The social and ethical consequences of a calculative mindset. *Organizational Behavior and Human Decision Processes* 125 (1): 39–49.

Zhong, C.-B. 2011. The ethical dangers of deliberative decision making. *Administrative Science Quarterly* 56 (1): 1–25.