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] and Anthony I. Jack[1,2]

[1]Case Western Reserve University, Department of Philosophy
[2]Case Western Reserve University, Inamori International Center for Ethics and Excellence

DRAFT: FOR ONLINE MINDS CONFERENCE ONLY. DO NOT FURTHER CIRCULATE OR CITE

1. Introduction

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…"* (Immanuel Kant, Critique of Pure Reason, 1787/1999, 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 ever reflective man, will have it and, for want of a recognized standard, will shape it for himself after his own pattern. (Immanuel Kant, Prolegomena, 367)

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 for the greater good of many? That these questions have remained at the heart of philosophy for so long, and that their 'solutions' (e.g., monism vs. dualism) seem to be incommensurable with each other, strikes us as enigmatic. Might the intractable nature of these and other appropriately identified problems reflect something peculiar about *us* rather than something peculiar about the way the world *is*? More specifically, might we instantiate competing cognitive processes which render these three philosophical problems problematic? Suppose that not only is this so, but also that

these cognitive processes arise from an unalterable feature of our biology. If so, then our neural architecture might prevent an ultimate solution that decides in favor of one competing view over another.

We think this is the case, and we seek to combine philosophical theory with empirical evidence to advance this account. This account maintains that the difficulties reconciling markedly different philosophical responses to these three questions arise from an unavoidable tension between two anatomically independent and functionally inhibitory neural networks, both of which are essential to human understanding. This account is motivated by the observation that both philosophers and non-philosophers experience difficulty in reconciling competing responses to these questions.

Because our account uses empirical evidence to address philosophical questions, it falls under the broad category of work in experimental philosophy. Because these differences are present in both philosophers and the folk, we do not privilege the viewpoints or judgments of one group over the other (e.g., philosophers vs. non-philosophers) in order to adjudicate metaphysical disputes and thereby declare one side victorious and the other mistaken. We believe the divergence found both between and within both philosophers and the folk is indicative of something much deeper than differences in philosophical education. Hence, we are interested first in establishing what the presence *of* and judgments *about* these perennial problems, including their associated phenomena, might reveal about the thinker. We deliberately delay addressing the question of what these judgments might reveal about the problems and phenomena themselves, pending a better understanding of the psychology of those judgments.

Our hypothesis is that different responses to the three philosophical questions of interest are driven by cognitive processes arising from two large-scale cortical networks. One of these networks, the Default Mode Network (DMN), is consistently activated by tasks that require connecting with and understanding one's own and others emotions and mental states (Amodio & Frith, 2006; Bzdok et al, 2012; Jack et al., 2013a; Jack, Dawson & Norr, 2013b; Li, Mai & Lieu, 2014; Reniers et al., 2012; Schilbach, Eickhoff, Rotarska-Jagiela, Fink & Vogeley, 2008; Van Overwalle, 2009; Van Overwalle, 2011; Van Overwalle & Baetens, 2009). The other network, the Task-Positive Network (TPN), is consistently activated by non-social tasks that require sustained attention, including mathematical, logical, mechanical and physical reasoning **(**Corbetta et al., 1998; Duncan & Owen, 2000; Goel, 2007; Jack et al., 2013a; Martin & Weisberg, 2006; Shulman et al., 1997). These two networks have a relationship such that engaging one network tends to suppress the other network (Fox et al., 2005; Uddin et al., 2009; Vincent et al., 2008**)**. This 'anti-correlated relationship' has been observed when humans are at rest, engaged in spontaneous thinking, in the brain scanner (Fox et al., 2005**)** and even in non-human mammals (Mantini et al., 2011) while anesthetized (Vincent et al., 2007). These findings, as well as mathematical models of the brain, suggest there are evolutionarily endowed neural constraints on cognition; in other words, the relationship between these networks reflects something about the brain's evolved network architecture and functionality, independent of any experimental demands and cultural influences.

We have advanced an account, dubbed the opposing-domains hypothesis, (Jack & Robbins, 2012; Jack et al., 2013a, 2013b; Jack, 2014; Jack, Robbins, Friedman & Meyers, 2014; Jack, Boyatzis, Nolan & Friedman, submitted; Robbins & Jack, 2006) according to which simultaneously activating core regions of the DMN above baseline while deactivating core regions of the TPN below baseline reflects various forms of emotionally engaged and other-oriented social cognition (the Phenomenal stance). In contrast, activating core regions of the TPN above baseline while simultaneously deactivating core regions of the DMN below baseline corresponds with various forms of analytic and scientific reasoning (including the Physical stance). Although these networks usually suppress each other, performance on some classes of tasks does involve coactivation of both DMN and TPN regions above baseline. On our account, this includes both creative thinking and various forms of emotionally disengaged and instrumental forms of self-serving social cognition (the Intentional stance). See Figure 1 for a schematic depiction of the model.

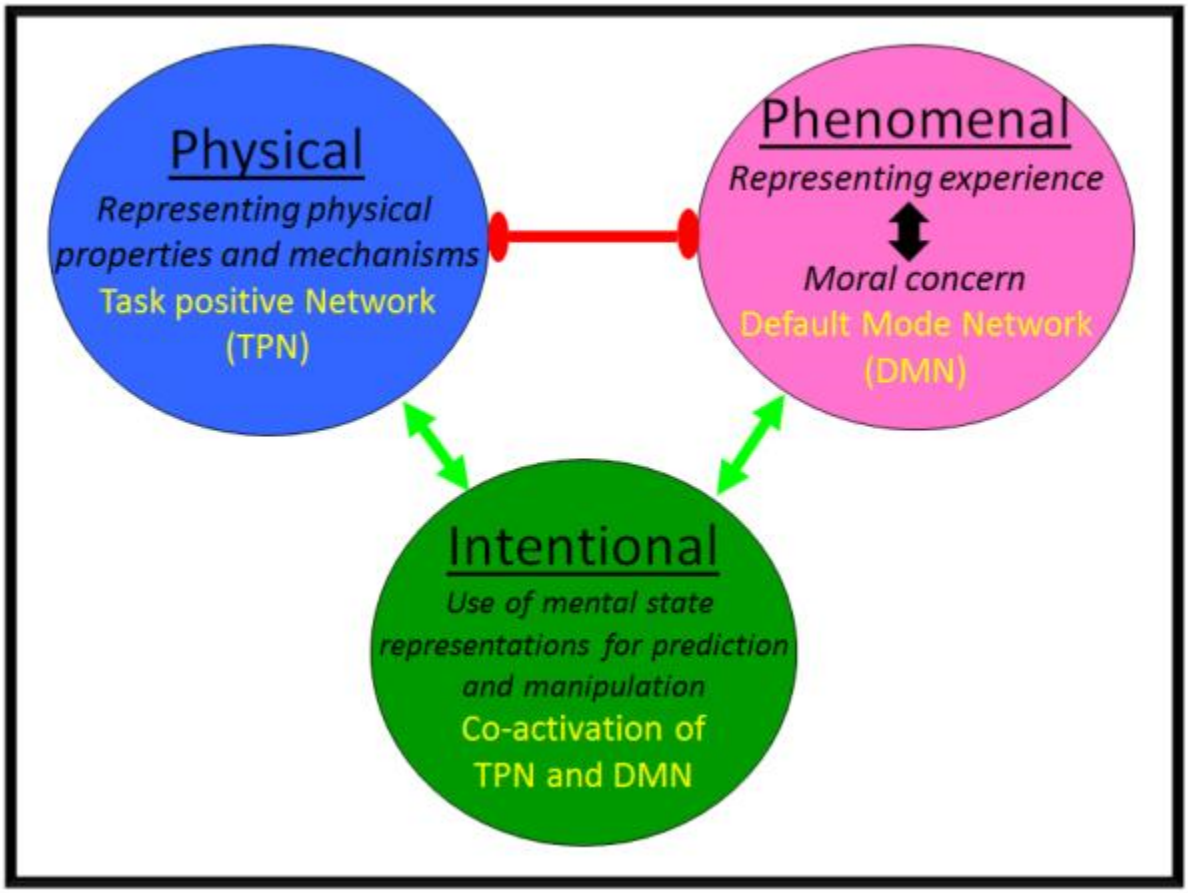

Figure 1: Three cognitive stances, their relationships to each other, and the brain networks involved. Bidirectional arrows indicate mutual compatibility; barbell indicates mutual antagonism. The Intentional stance is a distinct cognitive mode in which processes associated with the task positive network operate on representations stored in the default network. Nonetheless, there remains a fundamental tension between cognitive modes involved

in the representation of experiential mental states and certain moral concerns (the Phenomenal stance) and the representation of physical mechanisms (the Physical stance). The opposing-domains hypothesis holds that this tension represents the cognitive basis of the reciprocal inhibition between the default mode and task-positive networks, which in turn underlies the inability to fully bridge fundamentally different conceptual representations into a single and unified worldview.

The present paper is primarily concerned with articulating how the fundamental tension between the Phenomenal and Physical stances - and their underlying neural signatures - relates to the seemingly insoluble nature of the three problems mentioned above. In what follows, we advance an account attempting to explain why, despite millennia of philosophical analysis and centuries of empirical advances, there is no single and unified solution to any of these three problems; that is, a solution which would leave proponents of competing camps, each of which desires to solve the problem by an allegiance to their own agenda, equally satisfied (e.g., a solution which would definitively favor monism [or dualism] in a manner that addresses the concerns of dualists [or monists].

We believe the relationship between these two neural networks, and the distinct types of cognition they instantiate, is 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 viewpoint. In light of this, we don't share the presupposition which appears to guide so much work in philosophy, namely that the divergent worldviews between folk and philosophers – and even those evidenced among expert philosophers themselves – can be harmoniously reconciled within an ultimately unified and naturalistic framework. Instead, we believe that the problems themselves, and their competing solutions, emerge from an evolved neural architecture which imposes inevitable contradictions on our understanding of the world.

This idea that opposing philosophical beliefs are associated with different cognitive faculties is not novel. What is novel to our philosophically motivated and empirically constrained approach is that it is able to formulate falsifiable hypotheses that are susceptible to the experimental methods of cognitive science, the results of which have a direct bearing on philosophical theorizing. Others have advanced programs along similar lines (e.g., Cushman & Greene, 2012). Our project advances their work by using a novel and empirically grounded theory of cognition to derive testable hypotheses about two opposing domains of cognition and how their competition might underlie three historically intractable philosophical problems.

Since we do not privilege any one stance as providing ultimate insight into the truth, we demarcate ourselves from other projects seeking to 'explain away' metaphysically peculiar beliefs espoused by both the folk and philosophers (e.g., beliefs that are at odds with a naturalistic worldview, including intuitive dualism and intuitive libertarianism) as the product of

epistemically defective psychological processes (Greene, 2011; Nichols, 2014), intuitive biases (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012), cleverly designed intuition pumps (Dennett, 1984; 1993; 2013)[1], or philosophical ignorance (Devitt, 2012**)**. We further demarcate ourselves from projects investigating folk and/or philosopher's intuitions in order to shift the burden of proof – either directly or indirectly – between competing philosophical theses (e.g., Sytsma & Machery, 2010; unpublished). We eschew the foregoing views because our sense is that they fail to offer an empirically adequate account *of* the divergent views, and in turn they neglect to offer a suitably respectful account of why such metaphysically peculiar (e.g., anti-physicalist) worldviews persist under the reflective scrutiny of expert philosophers (e.g., Chalmers, 1997; Kripke, 1972; Nagel, 1974). We take this divergence in expert philosophical opinion seriously, and our theory and experiments are designed to provide novel insights illuminating both the origin *of* perennial problems and the perception *that* perennial problems are in fact problematic[2].

Utilizing experimental methods to test novel, falsifiable, and theoretically driven hypotheses concerning the cognitive origin of philosophical problems and their related beliefs does not, by itself, render a project anti-philosophical or spawn a new breed of (meta)philosophy. Despite our focus on cognitive architecture, we maintain that our experimental project shares a greater affinity with philosophically influential variants of conceptual analysis – analyses which served to illuminate the disparity between, for instance, the conceptual viewpoints or attitudes underlying the representation of phenomenal consciousness and physical mechanism, or between moral appraisal and mechanistic determinism – than it does with 'data driven' cognitive science[3]

---

[1] To be sure, we are not arguing that intuitive hunches and intuition pumps don't exist or fail to influence philosophical judgments. Rather, we want to highlight a critical difference between deliberately pumping intuitions with thought experiments, and intuitions which are themselves pumped by our neural architecture.

[2] When it comes to perennial philosophical problems, we regard the 'origin *of*' the problem, and the 'perception *that*' the problem is 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 (Dennett, 1993, 2013; Churchland, 1981). Philosophers espousing this view indeed acknowledge that there is a (temporarily) genuine and (temporarily) unavoidable 'origin of' the problem – namely, scientific ignorance. Consequently, a mature neuroscience should dissolve the 'perception *that*' the problem is intractable; time and scientific progress are the obstacles, not endogenous neural constraints. Others hold that philosophers and scientists are 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 (Sytsma & Machery, 2010; see especially Machery, 2013). Those subscribing to this view deny 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 phenomenal consciousness is traced to modern philosophers (e.g., Descartes, 1641/1996; Leibniz 1714/1992), not our endogenous neural constraints. The foregoing treatments are deflationary accounts of what we, like Kant (see section titled *Antinomies and neural structure*), view as deeply meaningful problems that have not only perplexed philosophers for millennia, but are bound to persist despite advances in scientific and philosophical understanding. We take this view because the emerging evidence suggests they are grounded in the functional architecture of the brain.

[3] Cognitive science, as we see it, seeks to systematically reveal the cognitive structure of the brain. Philosophy, on the other hand, seeks to resolve, gain insight into, or acquire meaning from certain problems and questions which, unbeknownst to us until recently, seem to arise *from* our cognitive structure. That the two endeavors are connected in this way is, one must admit, both auspicious and illuminating. Auspicious because studying the structure of the

(c.f. Knobe, forthcoming). We maintain that our project has philosophical import in a manner largely parallel with, yet still complementary to, these traditional methods of philosophical analysis. Consequently, we disagree that an entirely novel metaphilosophical framework must be constructed in order to fully appreciate the philosophical significance of the evidence revealed by our project. This is, of course, because our project does indeed aim to provide "the same sort of thing" (c.f. Knobe, forthcoming) that the particular projects we share an affinity with aimed to provide – namely, a diagnosis of the origin of incommensurable philosophical worldviews and their underlying conceptual attitudes (see sections 3.1 and 3.2). We simply take this approach one step further: In addition to investigating the concepts, worldviews and beliefs, we also investigating the underlying brain areas and cognitive processes which drive them, as well as the relationship between these brain areas and cognitive processes.

Now, if it can be demonstrated that our cognitive architecture renders certain philosophical questions intractable and problematic, shouldn't such a finding compel philosophers to refine and re-conceptualize the nature of these problems themselves (e.g., Kuhn, 1962/2012; also see Kant, 1787/1999)? We think so. And we are not alone in thinking this. We join the company of other philosophers who embodied this position, or whose work demonstrates an original attempt to elucidate a very similar point, including William James, Immanuel Kant, Thomas Nagel, and Peter Strawson.

The rest of the paper is laid out as follows. In Section 2 we briefly discuss the status of philosophical intuitions and their relevance to our project. In Section 3 we draw similarities between the philosophical import afforded by our experimental philosophical project and other more traditional, yet particular, philosophically significant species of conceptual analysis. In Section 4 we articulate in more detail the theoretical motivations, philosophical and empirical significance of our project.

2. Philosophical intuitions

There is much disagreement concerning the nature of intuition and whether intuitions, or other forms of more reflective reasoning, bear greater significance to philosophy. Some have argued that intuitions are not philosophically significant in any sense (e.g., Cappellen, 2012). Others have pointed out that we should evaluate their legitimacy on a case-by-case basis (e.g., Bengson, 2013). And others have argued that some of our intuitions are unreliable because they are unstable and influenced by arguably irrelevant factors (e.g., Machery, Mallon, Nichols & Stitch, 2004; Nichols & Knobe, 2007; Swain, Alexander & Weinberg, 2008; Weigmann, Okan & Nagel, 2012). Those debates are interesting in their own right and certainly serve to identify any methodological shortcomings associated with projects interested in intuitions.

However, our approach to experimental philosophy does not hinge upon investigating only certain kinds of responses (e.g., reflective judgments vs. immediate intuitions) from only

---

brain through a philosophical lens can reveal otherwise invisible philosophical insights; and illuminating because such insights can advance philosophical discourse in a traditional and philosophically consequential manner.

certain kinds of people (e.g., folk vs. philosophers). On the contrary, we see value in examining both philosophers and the folk, as well as various types of cognitive processes and judgments (e.g., psychological, behavioral, neurological) that are both directly and indirectly related to philosophical beliefs. In short, whether intuitions – whatever they turn out to be – are epistemically reliable or not is inconsequential for our argument here; we are not interested in intuitions exclusively.

We believe that arguing whether intuitions are 'the right sort of evidence' or are (not) immune to non-truth tracking factors is, in a sense, to focus on the trees instead of the forest. First, this is because we believe that certain philosophical problems are genuine and unavoidable phenomena arising from our neurology – they cannot be *fully* explained (away) by ambiguous text[4], not having minimal pairs, because of cultural platitudes, etc…, (e.g., Arico, 2010; Schwitzgebel & Cushman, 2015; Sytsma, 2013). We are not claiming that these factors do not influence people's intuitions, beliefs and judgments; we are merely claiming that these are superficial reasons that fail to illuminate the cognitive origin of irreconcilable philosophical beliefs and worldviews, especially those pertaining to our three problems of interest. (A rule of thumb for competing projects: If you can't explain the divergence in *both* the folk and expert philosophers, then you aren't even aiming at an account along the lines we are aimed at). Others have already provided evidence demonstrating the potential fecundity and explanatory afforded by such an approach. For instance, Schwizgebel & Cushman (2015) have shown that folk and philosophers are both susceptible to the same cognitive biases, providing evidence that examining both camps can reveal commonalities in our psychology which drive philosophical beliefs. Although the theoretical motivations and variables of interest behind our two projects differ, their work lends support to the general idea that philosophical positions may be rooted in basic cognitive processes – a hypothesis which can only be appropriately tested by studying both folk and philosophers.

Secondly, intuitions are not monolithic but are multifaceted and complex; they do not exist in isolation from our worldly experiences or psychology (e.g., Nado, 2014). Intuitions and other forms of judgment are influenced by individual differences in our psychology. This makes them especially interesting to study – especially as they relate to philosophical beliefs.

However, it is not enough to acquire a compendium of data about intuitions and personality differences without any theoretical motivations or predictions pertaining to the data. That isn't interesting. What is interesting are philosophically informed and empirically constrained theoretical predictions addressing why certain cognitive processes might influence certain philosophical beliefs and, moreover, the manner in which these cognitive processes relate to each other and their respective beliefs. Even better is if the experiments can be designed to

[4] Such problems as the 'Hard' problem of consciousness are *reflected* by our language, not *grounded* in it. Take the following footnote by Robert Nozick (1981), who emphasizes a philosophical practice that provides explanations for philosophical problems/disagreements instead of using coercive methods to force one's opponent to accept one's own view: "But the task is not the Wittgensteinan therapeutic one of removing or dissolving puzzles of confusions with linguistic roots, arising from mistaken analogies due to grammar, from "language on holiday", from thinking all terms are substantives and so forth. The explanatory problems are genuine and real…(p. 651-652, footnote 5).

test multiple competing hypotheses aiming to explain the same phenomena. This is precisely what our research agenda seeks to accomplish with the opposing-domains hypothesis.

However, this emphasis on experiments, intuitions and cognitive processes has led some to believe that whatever it is experimental philosophers are doing, it sure isn't philosophy. Or rather, that because experimental philosophy bears no resemblance to conceptual analysis, we must construct a novel metaphilosophical framework in order to fully appreciate its philosophical significance. Take the following quote by Joshua Knobe (forthcoming):

> "Experimental philosophy has sought to capture the patterns in people's intuitions through theories about underlying cognitive processes. In actual practice, this does not involve analyzing concepts, or doing something broadly similar to analyzing concepts, or engaging in some preparatory work that would eventually allow us to analyze concepts. It is not a matter of analyzing concepts at all; *it is something else entirely*." (our emphasis)

Knobe concludes that:

> "The vast majority of empirical research in experimental philosophy neither contributes nor attacks the conceptual analysis tradition…Experimental philosophy is pursuing philosophical questions in a way that is *genuinely new*, and the only way to properly explore its implications is to develop a correspondingly new metaphilosophical framework." (our emphasis)

We adopt a different perspective. It is not that we dispute Knobe's quantitative report (c.f., Knobe, 2015), but rather the idea that what experimental philosophy does need be starkly different from conceptual analysis. At least, the way we do experimental philosophy and the type of conceptual analysis afforded by our approach. Knobe suggests that a new metaphilosophical framework is required for interpreting the significance of experimental philosophy because theorists are examining the underlying psychological processes *instead* of using the intuitions to construct a theory of the concept investigated (e.g., using intuitions about free will and how the folk deploy the concept in order to arrive at a theory about the concept of free will). In contrast, we don't think a new metaphilosophical framework is required because examining the underlying psychological process is a complementary approach to using the intuitions to construct a theory of the concept investigated. Hence, we regard our approach to be both philosophically and empirically fecund, because it brings the two methods together in a manner that informs both disciplines.

It is important to emphasize that this difference in perspective partly results from the fact that we are not trying to interpret the significance of our results through the metaphilosophical framework utilized for the *particular* type of traditional conceptual analysis that Knobe discusses. For instance, in reference to some experimental work done on knowledge intuitions, Knobe writes that it "would not be accurate to say that work in this field is in any way providing

us with the *same sort of thing* that conceptual analysis had originally hoped to produce" (our emphasis). Knobe convincingly argues that most theorists do indeed try to explain an effect on people's intuitions by appealing to ordering effects or something of that nature, rather than contribute to the analysis of the concept. But this is neither our approach to experimental philosophy nor the type of conceptual analysis we have in mind. Knobe's analysis of experimental philosophy is therefore orthogonal to the motivations and methods of our project.

Our project does indeed aim to provide 'the same sort of thing' that philosophers like Immanuel Kant, Thomas Nagel, Peter Strawson and William James aimed to provide in their conceptual analyses – the elucidation of a tension "in us" (e.g. Strawson, 1974/2012, p. 67/25**)** between incommensurable philosophical viewpoints. Our project is continuous with and complementary to their analyses insofar as it provides empirical information pertaining to what it is, *in us,* that underlies the tension. At this point, we are in a better position to turn to Nagel and Strawson in order to both clarify the particular type of conceptual analysis we have in mind, and demonstrate that *that* type of conceptual analysis is in fact afforded by our approach.

3. From conceptual analysis to conceptual illumination.

It is generally agreed that conceptual analysis has made significant contributions to philosophical inquiry, especially among analytic philosophers (Jackson, 1998). Conceptual analysis involves an appeal to one's own (and others') intuitions in order to try and account for the independently necessary and jointly sufficient conditions that determine a concept, or for determining whether a concept can be said to accurately apply to its referent. There are distinct philosophical affordances associated with this form of conceptual analysis. For example, it enables philosophers to construct a theory of the concept *x*, which itself offers distinct predictions about its deployment and accuracy. Other forms of conceptual analysis focus on the relationship between two or more concepts, or perhaps more aptly, between two or more conceptual viewpoints. These forms of analysis are not so much concerned with enumerating criteria for the correct application of a concept. Instead, given that *x* is the case, these analyses aim to supply a *descriptive* account of what might be going on *in us* – whether at the conceptual level or faculty level – that might explain *x*. For example, given our inability to reconcile two disparate conceptual viewpoints, what might explain this inability? Or, given that something essential is left out from one solution to a problem, while something dubious is introduced by its competing solution, what might explain this tension?

Our approach to experimental philosophy shares an affinity with this particular species of conceptual analysis, which is best exemplified in the work of philosophers like Thomas Nagel (1974), Peter Strawson (1974/2012), Immanuel Kant (1787/1999) and William James (1906/1995**)**. These philosophers were not primarily interested in developing a theory of the concept *x* or determining whether concept *x* was accurately deployed. Instead, they sought to illuminate the presence of a tension between disparate and even opposing concepts or attitudinal viewpoints which are used to represent the same thing (e.g., for Nagel, the phenomenal nature of consciousness *and* the physical nature of consciousness), or to conceptualize an issue (e.g., for

Strawson, the significance that the truth of determinism might have for moral responsibility). Immanuel Kant is especially germane to this essay insofar as he articulated the possibility that certain perennial problems (e.g., antinomies) were grounded in the functional relationship between our (transcendental) cognitive faculties. Soon thereafter, William James went a step further and suggested that opposed philosophical worldviews may actually emerge from personality differences. In his own words, "[T]he history of philosophy is to a great extent that of a certain clash of human temperaments" (1906/1995, p.2). We shall have something to say about each of these philosophers in turn, especially as they relate to our project.

3.1 Nagel, the opposing-domains hypothesis, and the problem of consciousness

In *What is it like to be a bat?* Nagel didn't *do* traditional conceptual analysis, in which we appeal to intuitions to try and account for the independently necessary and jointly sufficient conditions that determine a concept. Instead, he 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 hinted[5], 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 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 which 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

[5] See footnote 11 in *What it is like to be a bat?*

> 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, and so he neither offered a solution to the mind-body problem[6] 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. As we mentioned above, and will further elaborate on below, emerging evidence is beginning to suggest that we are neurologically constrained from ever realizing such a framework. This is because what Nagel described as the subjective viewpoint shares an affinity with what we have dubbed the Phenomenal stance (see Robbins & Jack, 2006; also see section below titled *Opposing-domains theory - a distinct approach to X-Phi);* and what Nagel described as the objective viewpoint shares an affinity with Dennett's Physical stance (Dennett, 1989), a stance which is featured in our philosophical theory of cognition.

At its core, the Phenomenal stance instantiates the ability to empathize with and understand others', as well as one's own, subjective experiences; this requires activating core regions of the DMN above baseline and deactivating core regions of the TPN below baseline (Jack et al., 2013a; Jack et al., 2013b). And it its core, the Physical stance instantiates different forms of analytic and scientific reasoning; optimal performance on these skills requires activating core regions of the TPN above baseline while deactivating core regions of the DMN below baseline (e.g., Anticevic, Repovs, Shulman & Barch, 2010; Fassbender et al., 2009; Jack et al., 2013a; Shulman et al., 1997). In short, the relationship between these two networks can be thought of as a neural see-saw that oscillates between the two stances – and just like a real see-saw, neither side (i.e., neural network) can be fully engaged at the same time as the other.

Hence, the neurological tension underlying these two incommensurable modes of understanding (i.e., the Phenomenal stance vs. the Physical stance, respectively) 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 "water is H_2O" is completed within the same system, that is, by adopting the Physical stance and engaging its underlying neural hardware.

---

[6] Apart from the sketch he offers in his footnote 11 of *What it is like to be a bat?*.

In our view, as opposed to that of Jaegwon Kim (1984, 2005), 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 so vividly articulates, and Chalmers (1997) repeats, it just *seems* that something is left out from this translation. Our project 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 (i.e., the Phenomenal stance) transcends explicability in terms of the other domain (i.e., the Physical stance). The harmonious communication required for successfully translating between these domains is not fully realizable.

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. In other words, the disconnect *in us* between different modes of understanding and conceptual representation, the nature of which poses an obstacle to seamless conceptual integration.

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, 1787/1999, A424/B452). That is, he 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 itself becomes no less genuine, but the goal of arriving at a definitive solution in favor of one view over another certainly looks less sensible.

3.2 Strawson, the opposing-domains hypothesis, and the problem of free will

In *Freedom and Resentment,* Peter Strawson didn't *do* traditional conceptual analysis to argue that the problem of reconciling moral responsibility with determinism could be solved if only philosophers could deploy the same concept when referring to determinism[7]. Instead, he

[7] In fact, Strawson's treatment did not require a clear definition of determinism: "But how can I answer, or even pose, this question without knowing *exactly* what the thesis of determinism is? Well, there is one thing we do know; that if there is a coherent thesis of determinism, then there must be a sense of 'determined' such that, if that thesis is true, then all behaviour [sic] whatever is determined in that sense. Remembering this, we can consider at least what

clarified that the incommensurable positions pertaining to the dilemma of determinism – compatibilism vs. incompatibilism – emerged from a tension between different conceptual 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 attitudes vs. the objective viewpoint and its appeal to social policy/control). The viewpoint one adopted would in turn influence 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. 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, 1974/2012, emphasis in original, p. 76/33) and, out of reverence for these sentiments, puts forth dubious metaphysical propositions (i.e., claims at odds with a naturalistic worldview; 'panicky metaphysics', 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 but "are, profoundly, *opposed* to each other" (Strawson, 1974/2012, p. 66/24, emphasis in original). Indeed, in illuminating the tension between these opposing conceptual attitudes, Strawson noted that "what is above all interesting is the tension there is, *in us,* between the participant attitude and the objective attitude" (Strawson, 1974/2012, p. 67/25 our emphasis).

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. 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 and "understand[ing] how he [an agent] works" (Strawson, 1974/2012, p. 69/27). This viewpoint has a close affinity with Dennett's Intentional stance, which is also part of our philosophical theory of cognition. Indeed, for Strawson, this notion of 'understanding how he works' is not an understanding aimed at the 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. According to our theoretical framework, the Intentional stance is realized by a distinct blend of the cognitive capacities that underlie the Physical and Phenomenal stances. Hence, we believe the tensions

---

possibilities lie formally open; and then perhaps we shall see that the question can be answered *without* knowing exactly what the thesis of determinism is." (Strawson, emphasis in original, p.75/32)

illuminated by Nagel and Strawson correspond to cognitive divides that are not identical but do overlap[8].

Strawson didn't supply any empirical evidence in favor of his analysis. However, in support of Strawson's treatment is emerging evidence that recourse to the "panicky metaphysics of libertarianism" (Strawson, 1974/2012, p. 80/36) is driven by our moral sentiments unique to the Phenomenal stance. More specifically, 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, 1972/2012 p. 60/20) and endorse libertarianism (Jack, Friedman, Knobe & Luguri, in prep). By stepping into the Phenomenal stance, we suppress our tendency to adopt a socially disengaged and objective attitude, an attitude more indicative of the Intentional and Physical stances. This invites us to hold people responsible independent of any theoretical truths about the way the world is (Physical stance) – we are thus inclined to punish in the name of justice, for the sake of preserving a sense of moral order (Clark et al., 2014; Jack et al., in prep; Shariff et al., 2014).

As with Nagel, we might ask: If Strawson somehow had access to this evidence in 1974 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? Again, we disagree that a novel metaphilosophical framework should be constructed to fully appreciate the philosophical import of evidence which contributes to the 'same sort of thing' (c.f., Knobe, forthcoming) 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, 1787/1999). 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 practiced a distinct type of conceptual analysis. In fact, it seems more apt to refer to it as conceptual illumination. They neither aimed to provide a theory about concepts and their deployment nor declare that only one disputant was fundamentally mistaken. Instead, they sought to clarify *why* there was disagreement[9]. Their

[8] According to our account, the Intentional stance is distinct from the Phenomenal stance, but unlike the Physical stance is not in fundamental tension with it (see Fig. 1). Correspondingly, the problem of consciousness seems much more intractable than the dilemma of determinism. Hence compatibilism about phenomenal consciousness and physicalism is a less popular viewpoint than compatibilism about determinism and moral responsibility. Even explicit materialists who deny phenomenal consciousness endorse compatibilist views of determinism and moral responsibility (e.g., Dennett, 1984; 2013).

[9] Admittedly, Strawson did try to reconcile the two incompatible perspectives, but his attempt required 'radical modifications' by both parties. Whether Strawson thought he could successfully reconcile the two or not is irrelevant. What is important for the present purposes is that his analysis clarified that each conceptual attitude either leaves out something crucial (compatibilism leaves out moral sentiments) or introduces something dubious

treatments illuminated the presence of a tension between competing conceptual viewpoints, viewpoints which we impose on our own experience. We contend that this most closely resembles the philosophical contributions afforded by our approach to experimental philosophy. And, because we are aiming to provide 'the same sort of thing' that Nagel and Strawson were, why should we construct a novel meta-philosophical framework to appreciate the philosophical import of our project? The most salient difference between our project 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 philosophical 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 Antinomies and neural structure

Kant's project in the *Critique of Pure Reason* (1787/1999) 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. This is significant because any possible object of knowledge must conform to these *a priori* (transcendental) conditions. Some of these *a priori* conditions, however, inevitably 'overstep their bounds' and engender certain illusions.

One type of these illusions is an antinomy. Kant paid careful attention to the causes and origins of his antinomies,[10] which he described as 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. What are relevant are the two criteria which exemplify an antinomy for Kant. First, the competing propositions should be universally recognized across humans. In other words, we do not 'choose' to confront or construct these paradoxes – they simply emerge from the functional interplay of our cognitive faculties. Second, 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).

(incompatibilism introduces the 'panicky metaphysics of libertarianism'). Regardless of the details, it is the general approach that is significant for present purposes.

[10] see *The Transcendental Dialectic, Second Book, Second Chapter, The antinomy of pure reason* (CPR, 1787/1999, A405/B432-A567/B595).

We regard both the 'Hard problem of consciousness' and the dilemma of determinism as just such illusions, the source of which can be traced to the tension between the brain's default mode and task positive neural networks (DMN & TPN). And although not an 'illusion' itself, we have applied the opposing-domains hypothesis to the tension between deontological and utilitarian ethics (see below). For instance, belief in dualism is not only universal, with both philosophers and folk espousing dualistic worldviews (e.g., Berring, 2006; Bloom, 2009) (Kant's first criteria), but there is strong evidence in support of our hypothesis that this problem reflects a schism in our neurology, not our metaphysics. In other words, *that despite illuminating the source of the illusion the illusion will still persist* (Kant's second criteria).

It is worth clarifying that we are not endorsing metaphysical dualism (nor are we endorsing transcendental idealism in our reference to Kant and his antinomies). We are card carrying physicalists who see – and *feel* – the importance of seriously entertaining conceptual dualism.[11] We admire Kant because he was acute to the incommensurable 'Ideas' legislated by theoretical reason and practical reason. In his own inimitable way, Kant was able to articulate a great chasm between empirical truths and moral truths. He noted that the theses of his antinomies had a "certain practical interest" and served as the "cornerstones of morality and religion", while their antitheses "robs us of all these supports, or at least seems to rob us of them" (Kant, 1787/1999, A465/B493-A469/B497). In this way, Kant associated the thesis of each antinomy with moral excellence and the antithesis with a desire for attaining theoretical and empirical knowledge. This notion that there is a tension between empirical truths and moral truths, (i.e., between our understanding of the world and our place in it), is strongly supported by brain imaging evidence derived from tests of the opposing-domains hypothesis.[12]

3.4 Mapping the opposing-domains hypothesis onto the deontology vs. utilitarianism debate

We suggest that this philosophical and neural tension between understanding minds*, morally,* and understanding objects, *physically*, can provide some insight into the deontological vs. utilitarian debate, at least as it relates to certain types of moral judgments. 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 clarifies the cognitive underpinnings associated with the tendency to

[11] On our view, materialistic mapping of the brain can march forward, but not with the expectation of ever discovering a mechanism of consciousness or even eliminating folk psychology. A fuller understanding of the universe and our place in it can emerge by respecting the disparity between these two opposing viewpoints (TPN & DMN; the Physical stance and Phenomenal stance, respectively) and coupling systematic phenomenological reports with mechanistic investigations (e.g., Jack & Roepstroff, 2002).

[12] Of course, our theory was designed independent of Kant's philosophy. That we characterize the Phenomenal stance as a reflective and deliberate form of moral reason is not to be consistent with Kant but because of neuroimaging evidence (e.g., the default mode network is highly evolved [Jack, Robbins, Friedman & Meyers, 2014]]and susceptible to cognitive load [e.g., Rameson, Morelli & Lieberman, 2012; Meyer, Taylor & Leiberman, 2015]).

endorse deontological or utilitarian judgments in a particular type of moral dilemma – hypothetical footbridge dilemmas that aim to pit these opposing ethical programs against each other.

Ethical thinking has long been dominated by these two opposing systems attempting to guide moral action. According to deontological thinking, which is best 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 (Kant, 1785/2005; Johnson, 2014). 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 prosocial sentiments realized by the Phenomenal stance (see Jack et al., 2014 for a discussion).

According to utilitarian thinking, which is best exemplified by the work of John Stuart Mill, the right moral action is that which maximizes the aggregate happiness (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 that is not only absent from, but universally discouraged by, deontological thinking, especially when one understands the categorical imperative in terms of the Humanity Formula.

Consequently, it is our view that utilitarian thinking in these footbridge type moral dilemmas is facilitated by a natural inclination to adopt both the Intentional stance and Physical stance, with the latter having a stronger effect. In contrast, we believe that deontological judgments in these moral dilemmas are associated with an inclination to step into the Phenomenal stance, and thereby connect with and respect the humanity in others. If this hypothesis is correct, then only a particular type of emotional thinking should relate to deontological responses – namely concern relating to the moral patiency of others.

Our account also makes three other predictions. 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 are controlled and emerge from the DMN (e.g., Lindquist, Wager, Kober, Bliss-Moreau & Barrett, 2012; Rameson, Morelli & Lieberman, 2012). It is these latter sentiments in particular which, we submit, facilitate with connecting to other humans, as opposed to primitive emotions. Second, emotionally disengaged measures of 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 (e.g., James & Blair, 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. We have shown that deontological responses are selectively predicted by prosocial sentiments, not measures of emotionality or those associated with the Intentional stance (Jack et al., 2014). Many other studies have shown that utilitarian thinking in these dilemmas is associated with psychopathic personality traits, including increased callous affect, aggression, and interpersonal manipulative skills (Koenigs, Kruepke, Zeier & Newman, 2012; Jack et al., 2014; Patil & 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 (James & Blair, 1996). And in support of our hypothesis pertaining to the Physical stance, individual differences and experimental inductions of various forms of controlled reasoning (e.g., verbal, mathematical, analytic) increase utilitarian responses to these dilemmas (Paxton, Ungar & Greene, 2012) and increase antisocial behavior in general (Small, Lowenstein & Slovic, 2007; Wang, Zhong & Murnighan, 2014; Zhong, 2011;). Taken together, this work supports our view that treating humans as merely means to an end is facilitated by an inability 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 to "never treat humans as means to an end, but always ends-in-themselves" be associated with the prosocial sentiments instantiated by the DMN? As we have suggested elsewhere (Jack et al., 2014), we believe that psychological traits instantiated by the DMN, including empathy and compassion, increase the tendency to identify and connect with the humanity in others, and thereby reduce the tendency to view other humans 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 & Fiske, 2006; 2007; Harris, Lee, Capestany & Cohen, 2014; Jack et al, 2013b). Other researchers have shown that damage to key areas of the DMN involved with abstract emotional representation and emotional regulation (Roy, Shohamy & Wager, 2012) increases utilitarian responses to these dilemmas (Koenigs et al., 2012). And damage to brain areas that 'switch' between the DMN and TPN (e.g., Sridharan, Levitin & Menon, 2008; Menon & Uddin, 2010) results in a failure to fully engage the DMN during moral decision making, which corresponds with increased utilitarian responses to these dilemmas (Chiong et al., 2012).

We are not suggesting 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. As we have explained elsewhere, mature and flexible ethical thinking most likely emerges from the ability to efficiently oscillate between these stances and their underlying neural networks, a phenomenon which we believe allows individuals to trade-off competing ethical concerns (Jack et al., 2014; Boyatzis, Rochford & Jack, 2014). Moreover, our argument here is particular to a certain class of moral dilemmas involving certain sorts of agents, namely footbridge type dilemmas with humans. We have shown elsewhere that psychopathic

personality traits are not associated with utilitarian responses to these dilemmas when the agent in question is a non-human animal (e.g., dog; Jack et al., 2014).

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 his "Humanity Formula" of the categorical imperative, seems to be motivated by an appreciation for the sanctity of human life, an appreciation which we contend emerges from 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, Everett, Earp, Farias & Savulescu, 2014), it seems unlikely that treating humans as means to an end is driven by the desire to increase the aggregate happiness out of a sense of moral concern for those who would experience the aggregated increase.

4 Opposing-domains theory – a distinct approach to X-Phi

The previous sections argued that our approach to experimental philosophy, which is guided by the opposing-domains hypothesis, is continuous with and complementary to historically significant forms of philosophical analysis. In this section, we focus more on empirically testable hypotheses emerging from the opposing-domains hypothesis, and how we use them to illuminate the tension between competing philosophical worldviews.

According to our philosophically inspired model of cognition, there are at least three dissociable cognitive stances we can adopt to understand the world and our place in it. In addition to Dennett's Intentional and Physical stances, which cannot sufficiently account for our deeply entrenched prosocial and moral sentiments, Robbins & Jack (2006) introduced the Phenomenal stance. This is the stance we adopt when regarding an entity as a locus of moral concern. The Phenomenal stance exists in tension with the Physical stance, but the Intentional stance reflects a blended mode of cognition between both the Phenomenal and Physical stances (see Fig. 1 above). That the Phenomenal stance exists in tension with the Physical stance is at the foundation of our theory. This grounds our hypothesis that certain sorts of moral cognition are inextricably coupled with the tendency to represent an entity's experiential mental states as fundamentally non-physical in nature (given the reciprocal suppression between these two stances and their underlying neural architecture).

We have since mapped the Physical and Phenomenal stances onto distinct and functionally inhibitory neural networks (Jack et al., 2013a). There is converging evidence that the reciprocal suppression between these networks is biological in origin (Fox et al., 2005; Mantini et al., 2011; Vincent et al., 2007). Consequently, it is not *just* that regarding an entity as a moral patient suppresses our tendency to think about its underlying mechanistic constitution (and vice versa); instead, this antagonism manifests itself even in the absence of any experimental demands. This provides a biological basis for the problem of consciousness.

The neural tension reflected by these three stances should be reflected by individual variation in the psychological components comprising each. For instance, we hold self-reported empathic concern, as measured by the the Interpersonal Reactivity Index - Empathic Concern (IRI-EC; Davis, 1983)[13] – or the absence of its converse, callous affect – (Self-Report Psychopathy, Callous Affect; SRP-CA [Paulhus, Neumann, & Hare, 2009]) – as the signature other-oriented emotion characteristic of the Phenomenal stance. Other moral sentiments, including retributivist justice **(**Jack et al., in prep**)** and intended prosocial sentiments (Jack et al., submitted), are also unique to the Phenomenal stance. Measures of analytic reasoning (Cognitive Reflection Test [CRT]; Frederick, 2005) and physical reasoning (Intuitive Physics Test [IPT]; Baron-Cohen et al, 2001) are better suited for the Physical stance. There are many well-validated measures for assessing one's ability to adopt the Intentional stance (e.g., Reading the Mind in the Eyes, Baron-Cohen et al., 2001; Diagnostic Analysis of Nonverbal; Accuracy [DANVA], Nowicki & Duke, 2001; Interpersonal Reactivity Index-Perspective Taking [IRI-PT], Davis, 1983). This is important because our theory makes novel and falsifiable predictions about the relationship between individual variation in these abilities and one's metaphysical and philosophical inclinations. We have already discussed some of these (above) in relation to the deontological vs. utilitarian ethics debate.

For instance, consider our prediction that psychopaths should have trouble perceiving *that* there is in fact a problem of consciousness (see Robbins & Jack, 2006, p. 76). According to our model, belief in dualism should be engendered by moral concern because genuine moral concern requires that we suppress thinking about an entity in terms of its underlying physical constitution. This is facilitated by higher levels of empathy. Hence, a natural inclination to step into the Phenomenal stance, via higher levels of empathy, should relate to belief in dualism.

Through a series of experiments we have demonstrated a robust negative relationship between callous affect, the signature diagnostic criteria of psychopathy, and belief in dualism (Jack, 2014). This relationship remains significant after controlling for gender, age, education level, religious belief, and performance on measures of analytic and physical reasoning. In fact, *dis*belief in dualism is best predicted by deficits in the Phenomenal stance (i.e., deficits in empathy) among those who scored highest on measures of analytical thinking. This stands in contrast with Dennett's view that beliefs which are at odds with a naturalistic world view should be curbed by a predilection for careful, analytic thought. We suggest that individuals with both high levels of analytic thinking and high levels of empathy are better able to appreciate the distinct affordances offered by these opposing cognitive stances. Consequently, they may be more adept at discerning when one cognitive stance is preferred to the other, and thus entertain the possibility that it is not irrational to hold conflicting beliefs when each belief is associated with an opposing domain of cognition (see conclusion for an elaboration on this point).

Now, at this point we draw your attention to the fact that such hypotheses – e.g., that psychopaths should be inclined toward a monistic worldview – serve to clarify that the tension

[13] The validity of using this measure has been supported by similar results using peer-report measures of empathy, as indexed by the Emotional and Social Competency Inventory (ESCI; Boyatzis & Goleman, 2007).

between such incompatible metaphysical theses as monism and dualism emerges from the functional relationship between the DMN and TPN. Our project thus appeals to judgments of both philosophers and the folk *as data* for testing philosophically consequential and empirically falsifiable hypotheses about the neural and psychological origins of judgments pertaining to these philosophical problems. We don't study folk (and expert) intuitions about philosophical phenomenon *x* (e.g., phenomenal states) in order to vindicate the veracity of philosophical phenomenon *y* (e.g., explanatory gap). Nor do we study intuitions about philosophical phenomenon *x* in order to weaken arguments pertaining to beliefs about philosophical phenomenon *y* (e.g., Sytsma & Machery, 2010). Rather, we investigate folk (and expert) judgments *about* philosophical phenomenon *x* in order to use such judgments *as data* for testing falsifiable hypotheses derived from a metaphysically neutral and neurologically grounded theory of cognition that predicts what drives (dis)belief in phenomenon *x*.

The opposing-domains theory is metaphysically neutral and neurologically grounded insofar as its claims are independent of what exists *out there* because it makes predictions about the neurological mechanisms and psychological faculties that incline people to one or another competing view pertaining to the philosophical phenomenon of interest. By this we mean that the theory itself would furnish the same predictions about the cognitive mechanisms inclining people to adopt, for example, a dualistic vs. monistic worldview, even if both of these metaphysical theses somehow turn out to be false. Unlike other projects, we do not presuppose the truth of a metaphysical thesis and let such a presupposition determine the value of folk and philosophers intuitions. We are not in the business of trying to harmoniously reconcile the divergent views between the folk and philosophers by *explaining away* the divergence. We believe there is sufficient evidence supporting the claim that we are cognitively constrained from realizing such a unified framework.

Once this is understood – that we appeal to data in order to understand *the origin* of divergent intuitions, worldviews, or beliefs, rather than evidence *in favor* of one intuition, worldview, or belief – the comparisons between consulting the folk on philosophical matters and consulting the folk on scientific matters becomes illegitimate. For instance, some have argued that because physicists aren't concerned with what the folk think about falling bodies *qua* falling bodies it follows that philosophers shouldn't be concerned with what the folk think about, for instance, phenomenal consciousness *qua* phenomenal consciousness (Papineau, 2011). On this view, folk judgments are philosophically inconsequential because philosophers ought to be concerned with the philosophical phenomena as they are in themselves (e.g., qualia, truth, freedom), not the way people think about the things themselves.[14] And, the argument goes, because the folk are unfamiliar with these phenomena, they could not possibly teach us anything about them (e.g., Machery, 2013).

[14] In *What is x phi good for* (2011), David Papineau writes "[W]hen philosophers study knowledge, consciousness, free will, moral value, and so on, their first concern is with these things themselves, rather than with what people think about them" (p.83).

But we are not interested in consulting the folk or philosophers exclusively to further understand what, for example, phenomenal consciousness *is*. Instead, we are interested in testing falsifiable hypotheses about the cognitive mechanisms inclining people to certain beliefs about phenomena such as phenomenal consciousness, free will, and what is good. In fact, we would go so far as to say that discussion of the biological basis of phenomenal consciousness (i.e., phenomenal consciousness as it is in itself), is to commit a category mistake (Ryle, 1949/2009). On our view, it is the insurmountable conceptual dissonance between the Physical and Phenomenal stances (and their underlying neural architecture) which gives rise to the feeling that there is in fact something like phenomenal consciousness, as it is in itself, which exists independent of our minds. The way that the phenomenon of consciousness is understood in one domain engenders the feeling that it must be wholly incommensurable with – antithetical to – the way it is understood in the other domain. By emphasizing one stance to the exclusion of the other, so that we bring that stance to bear on (the conception of) that phenomena which more properly lie in the other stance's domain, we come to commit such category mistakes. 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; the same goes for adopting the Phenomenal stance to try and understand the biological mechanisms of consciousness. We suggest that each stance has evolved to successfully navigate the world when adopted for certain representations in certain situations. Consequently, we should strive to balance our understanding of the world, and our place in it, from these opposing vantage points we impose on our own experience.

4.1 Epistemic credibility in X-Phi projects qua cognitive science

All but one of the measurements used in our laboratory – our 5-item dualism measure (Jack, 2014) – have been designed independent of our theory. The measurements we use for assessing the Phenomenal stance (e.g., IRI-EC; Davis, 1983; ESCI; Boyatzis & Goleman, 2007), Physical stance (CRT. Frederick, 2005; IPT, Baron-Cohen, 2001) and Intentional stance (Reading the Mind in the Eyes, Baron-Cohen et al., 2001; DANVA, Nowicki & Duke, 2001; IRI-PT, Davis, 1983) are all externally designed and well-validated for assessing distinct psychological capabilities. That our hypotheses have been supported by using these independently designed measures should, we believe, increase the epistemic credibility and truth seeking ability of the opposing-domains hypothesis. Indeed, there is something epistemically dubious about a program that seeks support for its own hypotheses by using measurements that were explicitly designed to assess *that* hypothesis. This raises some concerns, although perhaps many of which are wholly unavoidable in a novel scientific discipline (Kuhn, 1962/2012).

That other projects often design their own novel measurements with their hypotheses in mind reflects a deeper and more significant problem – positing the existence of philosophically motivated psychological processes for no reason other than to explain a counterintuitive effect on

philosophical judgments, or counterintuitive judgments themselves.[15] Such projects run the danger of failing to illuminate any genuine cognitive origins to the problems they address. Instead, at their worst, such projects may be merely presupposing the truth of a metaphysical thesis and explaining away any incompatible intuitions by appeal to purely invented cognitive processes. The flaw in this kind of approach is that it presupposes from the outset that any intuitions which don't align with the presupposed veridical thesis must be somehow defective. This presupposition compels the theorist to posit cognitive processes that can explain away the divergence or debunk the contradictory beliefs. However, if we (and Kant, James and Nagel and Strawson) are correct that certain philosophical problems arise as an inevitable result of our cognitive structure, then there is no good epistemic basis for positing cognitive processes for the purposes of explaining away the existence of contradictory beliefs. Any projects which actually fit this schema might be better described as poorly motivated and epistemically dubious science, rather than a type of philosophy. While this may appear an unfair caricature of X-Phi, there are projects that have this surface appearance and which therefore need some scrutiny.

Our main point is that the epistemic propriety of ad-hoc hypotheses put forth by such projects to account for counterintuitive results, or to explain away certain inconsistencies, is questionable because it is often the case that these unique processes are put forth to account for *that* particular philosophically relevant result. And, as Knobe (forthcoming) highlights, there is little reason to believe that we must posit a novel and strictly philosophically relevant psychological process to account for any counter-intuitive findings. But even if such a philosophically motivated process does gain traction, it will require systematic, interdisciplinary collaboration to vindicate the significance of the psychological process(es) in other cognitive domains, as well as isolate its (their) neural underpinnings.

The cognitive mechanisms and psychological processes intrinsic to our model are already recognized as such across disciplines. Hence we are able to offer testable hypotheses about how the cognitive processes integral to our theory not only relate to philosophical worldviews and judgments, but how their underlying processes relate to each other. By positing a fundamentally new process, the theorist is unable to explicate how that process might interact with other established psychological processes. Because the opposing-domains hypothesis is empirically constrained and informed by neuroscience, it does not face this difficulty. Furthermore, because we make testable claims about behavior, psychology, and neurology, our theory is falsifiable at several different experimental levels and thus many different experimental methods. Theories which posit a new process do not exhibit such 'vulnerability' to falsification as the opposing-domains hypothesis; we regard all of the foregoing as providing a promising epistemological edifice to our theory (e.g., Popper, 1962/2014).

[15] As we mentioned earlier, this is Knobe's diagnosis of what most experimental projects are doing: "It [X-Phi] consists of identifying surprising effects in people's intuitions and explaining those effects in terms of underlying cognitive processes," (forthcoming).

4.2 Epistemic credibility in X-Phi projects qua philosophy

In the previous section we questioned the epistemic credibility of many X-Phi projects insofar as they attempt to make claims about the nature of cognition. In essence, our argument was that many such projects will be motivated to make ad-hoc claims about cognitive processes to support their philosophical worldview, and that this is unlikely to clarify the relationship between philosophical worldviews and an empirically driven understanding of cognition.

A second issue concerns the epistemic credibility of standard X-Phi projects to provide evidence in support of particular philosophical claims or worldviews. This concern has been raised before by critics of X-Phi, and we find considerable accord with many of their concerns.

In what way could – or should – determining what the majority of the philosophically uneducated believe about philosophical phenomenon *x* be informative of the veracity of philosophical phenomenon *x* (or its related worldview)? That the scientifically uneducated folk are more likely to believe that an object will behave according to Aristotelian physics than Newtonian physics (McCloskey, 1983) is inconsequential to the truth of the matter. The emphasis adopted in many X-Phi projects on aggregating folk intuitions concerning various philosophical dilemmas, and then using such data to support or weaken a particular philosophical thesis seems inherently problematic to us.

For instance, regarding mereological composition, Rose and Schaeffer (2014) declare that because folk intuitions fail to align with expert intuitions "the folk deserve to be ignored" on this issue (p. 1). We have some sympathy for Rose and Schaeffer's response. For example, take the case of dualism vs. physicalism. We have the impression that, if it were to be found that most folk believe in physicalism, then some X-Phi philosopher would likely cite this evidence in favor of the truth of physicalism. However, most surveys (including but not limited to our own) indicate that the majority of the folk are dualists. Oddly, we do not know of any X-Phi philosopher who has cited this evidence as favoring dualism – presumably because X-Phi philosophers are strongly inclined to physicalism.

The alignment, or lack of alignment, between the philosophically uneducated and educated is no indicator of the truth. If both the philosophically uneducated and educated espouse the same worldviews, and if the philosophically educated (believe they) arrived at their conclusion through rigorous philosophical reflection, does it follow that the philosophically uneducated arrived at the same conclusion through the same philosophical reflection? Their lack of philosophical training would suggest otherwise. Alternatively, if the folk and philosophers tend, on average, to arrive at different conclusions (e.g., with regard to physicalism vs. dualism), we question whether the most opportune strategy is to proceed to attempt to identify a *source of error* in folk thinking.

Consequently, our emphasis is different from the many X-Phi projects that place an experimental emphasis on the mean or aggregated views of the folk. Instead, we focus on identifying the source of divergences which exists *within both* the folk and philosophers. For instance, philosophers are both dualists and physicalists, and ordinary folk similarly split between these two. That this divergence occurs in both groups is what interests us, regardless of

the fact that most of the folk are dualists and most (Western) philosophers are physicalists. That there is such divergence in both the philosophically educated and uneducated is suggestive of something interesting about minds in general, quite independent of the influence that philosophical education has upon them. Hence, what we attempt to do is explicate the tension between certain philosophical beliefs (e.g., physicalism vs. dualism) by recourse to our cognitive architecture rather than philosophical prowess. In other words, we aim to provide a philosophically external account that explains why someone might espouse dualistic beliefs (for example), rather than internally accounting for dualistic beliefs by recourse to Dualistic theories proper.

In closing this section, we want to reiterate that our interest in studying folk and philosophers intuitions about philosophical phenomena lies in testing theoretically motivated hypotheses, the results of which can illuminate the neurological and psychological processes underlying the cognitive origin of the philosophical phenomenon in question (or beliefs about the philosophical phenomenon in question). Our theory makes testable hypotheses about the relationship between cognitive processes and philosophical beliefs, as well as the relationship the implicated cognitive processes share with each other. It is in this way that the results of our project supply philosophically independent reasons that *explain why* some people are more inclined to one competing metaphysical thesis over its antithesis (or vice-versa). And again, the project is independent of any particular philosophical worldview because our theory grounds the origin of intractable problems in our cognitive constitution.

4.3 Why do philosophers disagree?

In the previous section we noted that a general problem for many X-Phi projects arises when divergence emerges between philosophers and the folk. When such discrepancies arise, some philosophers endorse the folk view whereas others dismiss it. Since there appears to be no powerful reason to endorse one strategy over the other, it is quite unclear what contribution queries about folk belief contribute to the resolution of philosophical disputes. Our approach is different because we focus on understanding divergences of views which exist in both camps. Neither group is privileged over the other. However, we focus on measuring these differences in the folk merely because they lack a philosophical education which might otherwise bias their beliefs – testing the folk allows a clearer and less encumbered view of how temperamental differences drive philosophical beliefs.

So, why *do* philosophers disagree? Do we, those who have been trained in logical reasoning and analysis of arguments, *really* believe that fundamental differences in philosophical belief merely result from minor differences in philosophical argumentation, different steps taken along the path to (T)ruth? Do we believe that once certain errors in our reasoning are identified, or the 'affect-laden' scenarios are seen for what they are independent of their 'emotional-pull', that prior disagreements will dissolve under the pressure of rational arguments that *force* their

conclusions upon us? Robert Nozick rather comically opposes this view when he writes that it would be "implausible" to suppose that the degree to which expert philosophers are committed to various premises and conclusions is "independent of how strongly he wants certain things to be true. The various means of control over conclusions explain why so few philosophers publish ones that (continue to) upset them" (Nozick, 1981, p. 2-3). Because we share an emphasis with Nozick on "finding harmony in apparent tensions and incompatibility" (Nozick, 1981, p. 10), while other experimental philosophers seem more concerned with using X-Phi data to support or weaken one of two (or more) opposing philosophical beliefs (e.g., dualism vs. monism), it is not surprising that such agendas have failed to utilize measures of individual differences. They have failed to take seriously that philosophical worldviews might arise from anything other than reason, or lack thereof. However, as mentioned above, Schwitzgebel & Cushman (2015) have shown that *both* expert and non-expert philosophers are susceptible to the same framing and ordering effects, suggesting that reasoning skills and philosophical education might be less important in sculpting our philosophical worldviews than previously supposed.

Perhaps both folk and philosophers are offering post-hoc rationalizations for philosophical convictions that are driven, at least in part, by temperamental factors (e.g., Haidt, 2000). That it does not appear to have more broadly occurred to X-Phi philosophers that their own beliefs may be driven by their temperament seems surprising when we consider, first, that so many of them are familiar with work in moral psychology that suggests this conclusion, and second, that influential thinkers throughout the history of philosophy have not only sought to identify the flaws of reason, but have even claimed that "reason is, and ought only to be the slave of the passions" (Hume 1738/1975). Consider the following quote by William James:

> "[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 1906/1995, p. 2)

Not only did James capture the idea that philosophers hold reason as superior to emotions or temperaments, and consequently – though for James, erroneously – as the guiding force of their arguments, but he suggests that our temperaments might be "the potentest [sic] of all our premises" (p.3). James goes 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 opposing-domains hypothesis is uncanny. 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. Admittedly James identifies tender-mindedness with idealism, whereas our work identifies it with dualism; however, both these philosophical viewpoints share the feature of privileging (experiential) mind over matter (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 (task positive network; TPN) and principles are aligned with the moral/ethical (default-mode network; DMN). Even if one disagrees with our drawing a parallel between James' division of facts and principles with the opposing-domains division of empirical/theoretical thought and moral/ethical sentiments, 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 worldview. Given James' status and influence on modern philosophy and psychology, it is rather odd that others have not paid attention to this insight and incorporated individual difference measures into their experimental projects as means of systematically testing hypotheses about the origins of philosophical beliefs.

In sum, the logic behind our own approach of emphasising 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 some individuals 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. In Kant's terminology, and a view echoed by James, individuals who strive for theoretical and empirical knowledge over and above a desire to cultivate moral excellence are more likely to endorse determinism, physicalism, utilitarianism, and reject belief in God **(**Jack et al., 2014; Jack et al., submitted; Jack et al., in prep)

5. Conclusion

We have argued that our experimental project shares a close affinity with classical forms of conceptual analysis in philosophy, the major difference being that our epistemic basis relies on data from the cognitive sciences, not our own introspection or logical analysis. We further argued that using these data complements these philosophically traditional analyses in ways that introspection and logical analysis alone could not possibly achieve. The type of conceptual 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, Masicampo, & DeWall, 2009; Vohs & 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 the inability to simultaneously deploy our neural networks underlying moral cognition 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 these and other philosophical problems. 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 and using them in the appropriate context (Phenomenal stance vs. Physical stance, and their underlying neural architecture).

This distinguishes our approach from other competing projects that (i) investigate psychological processes as evidence in support of a 'debunking' argument (e.g., Greene 2011; Nichols, 2014), or (ii) focus less on the psychological processes and more on the intuitions themselves, where the intention is to shift the burden of proof in philosophical arguments, either directly or indirectly (e.g., Machery, 2013 Sytsma & Machery, 2010; Sytsma & Machery, unpublished). These projects suffer from at least three serious shortcomings. First, they fail to seriously account for the divergent philosophical worldviews found *within* expert philosophers. Second, they run the risk of inaccurately characterizing the psychological processes underlying opposing philosophical worldviews, and thus the philosophical worldviews themselves, which can in turn cause us to lose sight of the humanity in individuals who adhere to either of the competing views (e.g., see Jack et al., 2014 for a discussion on this point). Third, by advancing a program that is not motivated by a philosophically neutral and empirically grounded theory of cognition, they are likely to posit unscientific ad-hoc hypotheses to account for counterintuitive and/or discrepant results.

Finally, we want to highlight two different ways in which beliefs and attitudes may be judged (ir)rational and how one of these relates 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 this latter sense it would be irrational for you to believe that you are not reading this essay right now because you are in fact reading this essay (hopefully). It is also in this latter sense that many philosophers mistakenly believe that only one of two (or more) ostensibly incommensurable viewpoints *must* be correct – that is, holding both of them simultaneously seems irrational in this sense. We have gone to great lengths to explain why we disagree: Upon learning that our own cognitive architecture predisposes us to treat certain beliefs and attitudes as contradictory, and thus irrational in this later sense – simply because they emerge from distinct and profoundly opposed cognitive processes – we should rethink the decision to point and laugh at those who are irrational in this sense.

Acknowledgements:

We would like to thank Chris Haufe, Joshua Knobe and Phillip Robbins for their critical and helpful comments on an earlier version of this manuscript.

WORKS CITED

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

Anticevic, A., Repovs, G., Shulman, G. L., & Barch, D. M. (2010). When less is more: TPJ and default network deactivation during encoding predicts working memory performance. *Neuroimage*, *49*(3), 2638-2648.

Arico, A. (2010). Folk psychology, consciousness, and context effects. *Review of Philosophy and Psychology*, *1*(3), 371-393.

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

Baron□Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). 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.

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

Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research*, *86*(3), 495-532.

Bering, J. M. (2006). The folk psychology of souls. *Behavioral and brain sciences*, *29*(05), 453-462.

Bloom, P. (2009). *Descartes' baby: How the science of child development explains what makes us human*. Basic Books.

Boyatzis, R.E. & Goleman, D. (2007). Emotional and Social Competencies Inventory, Boston: The Hay Group.

Boyatzis, R. E., Rochford, K., & Jack, A. I. (2014). Antagonistic neural networks underlying differentiated leadership roles. *Frontiers in human neuroscience*, *8*.

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

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

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

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.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 67-90.

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

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

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience*, *7*(3), 269-279.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*,*44*(1), 113.

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

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

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

Descartes, R. (1641/1996). Discourse on Method and Meditations, trans. *Laurence J. Lafleur. Indianapolis: Bobbs–Merril Educational Publishing*.

Devitt, M. (2012). The role of intuitions in the philosophy of language. In G. Russell & D. Graff (Eds.), *In the Routledge Companion to the Philosophy of Language* (pp. 554-65). New York, NY: Routledge.

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

Fassbender, C., Zhang, H., Buzy, W. M., Cortes, C. R., Mizuiri, D., Beckett, L., & Schweitzer, J. B. (2009). A lack of default network suppression is linked to increased distractibility in ADHD. *Brain research*, *1273*, 114-128.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 25-42.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (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.

Fox M. D., Corbetta M., Snyder A. Z., Vincent J. L., Raichle M. E. (2006). Spontaneous

neuronal activity distinguishes human dorsal and ventral attention systems. *Proc. Natl. Acad. Sci. U.S.A.* 103 10046–10051.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in cognitive sciences*,*11*(10), 435-441.

Greene, J. D. (2011). Social neuroscience and the soul's last stand. *Social neuroscience: Toward understanding the underpinnings of the social mind*, 263-273.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low neuroimaging responses to extreme out-groups. *Psychological science*, *17*(10), 847-853.

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social cognitive and affective neuroscience*,*2*(1), 45-51.

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

Hume, D., *(1738). A Treatise of Human Nature*, edited by L. A. Selby-Bigge, 2nd ed. revised by P. H. Nidditch, Oxford: Clarendon Press, 1975.

Hyatt, C. J., Calhoun, V. D., Pearlson, G. D., & Assaf, M. (2015). Specific default mode subnetworks support mentalizing as revealed through opposing network recruitment by social and semantic FMRI tasks. *Human brain mapping*.

Jack, A. I. (2014). Conceptual Dualism. *Oxford Studies in Experimental Philosophy*, *1*, 5.

Jack, A.I., Boyatzis, R.E., Nolan, S.N., & Friedman, J.P., (submitted). Why Do You believe in God? Opposing Relationship of Empathy and Analytic Thinking

Jack, A. I., & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, *3*(3), 383-403.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: from stimulus–response to script–report. *Trends in cognitive sciences*, *6*(8), 333-339.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013a). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385-401.

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

Jack, A.I., Friedman, J.P., Knobe, J., & Lugrui, J. (in prep). Moral sentiments drive the belief that mind and body are distinct.

Jack, A. I., Robbins, P. A., Friedman, J. P., & Meyers, C. D. (2014). More than a feeling: counterintuitive effects of compassion on moral judgment. *Advances in Experimental Philosophy of Mind*, 102-125.

Jackson, F. (1998). From metaphysics to ethics.

James, W. (1906/1995). *Pragmatism*. Courier Corporation.

James, R., & Blair, R. (1996). Brief report: Morality in the autistic child. *Journal of autism and developmental disorders*, *26*(5), 571-579.

Johnson, Robert, "Kant's Moral Philosophy", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/kant-moral/>.

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

Kant, I. (1785/2005). Groundwork of the Metaphysic of Morals (T. Abbott, Trans.).*Ontario: Broadview Press.(Original work published 1785)*.

Kant, I., Guyer, P., & Wood, A. W. (Eds.). (1787/1999). *Critique of pure reason*. Cambridge University Press.

Kant, I. (1902). *Prolegomena to any future metaphysics that can qualify as a science* (No. 53). Open Court Publishing.

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

Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.

Knobe, J. (2015). Philosophers are doing something different now: Quantitative data. *Cognition*, *135*, 36-38.

Knobe, J. (forthcoming). Experimental philosophy is cognitive science. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (1st ed., pp. Xx-xx). Blackwell.

Koenigs M., Kruepke M., Zeier J., Newman J. P. (2012). Utilitarian moral judgment in psychopathy. *Soc. Cogn. Affect. Neurosci.* 7 708–714 10.1093/scan/nsr048

Kripke, S. A. (1972). *Naming and necessity* (pp. 253-355). Springer Netherlands.

Kubit B., Jack A. I. (2013). Rethinking the role of the rTPJ in attention and social cognition in light of the opposing domains hypothesis: findings from an ALE-based meta-analysis and resting-state functional connectivity. *Front. Hum. Neurosci.* 7:323. 0.3389/fnhum.2013.00323

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago press.

Leibniz, G. W. (1714/1992). *Discourse on Metaphysics and the Monadology* (trans. George R. Montgomery).

Li, W., Mai, X., & Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us. *Frontiers in human neuroscience*, *8*.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, *35*(03), 121-143.

Machery, E.(2013). Interviewed by Tony Sobrado. *Edouard Machery & Tony Sobrado on Consciousness*. PodBean podcast. http://tonysobrado.podbean.com/2013/12/08/edouard-machery-tony-sobrado-on-consciousness/

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1-B12.

Mantini, D., Gerits, A., Nelissen, K., Durand, J. B., Joly, O., Simone, L., ... & Vanduffel, W. (2011). Default mode of brain function in monkeys. *The Journal of Neuroscience*, *31*(36), 12954-12962.

McCloskey, M. (1983). Intuitive physics. *Scientific american*, *248*(4), 122-130.

Martin, A., & Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cognitive Neuropsychology*, *20*(3-6), 575-587.

McKenna, M., & Russell, M. P. (Eds.). (2012). *Free will and reactive attitudes: perspectives on PF Strawson's" Freedom and resentment"*. Ashgate Publishing, Ltd..

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

Meyer, M. L., Taylor, S. E., & Lieberman, M. D. (2015). Social working memory and its distinctive link to social cognitive ability: An fMRI study. *Social cognitive and affective neuroscience*, nsv065.

Mill, J. S. (1998). Utilitarianism (R. Crisp, Ed.).

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

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, 435-450.

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

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*(4), 663-685.

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*,*2011*(52), 83-88.

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in psychology*, *5*.

Paulhus, D. L., Neumann, C. S., & Hare, R. D. (2009). Manual for the self-report psychopathy scale. *Toronto: Multi-health systems*.

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

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

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

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

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

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *Neuroimage*, *41*(4), 1447-1461.

Rose, D., & Schaffer, J. (2014). Folk mereology is teleological.

Roy M., Shohamy D., Wager T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends Cogn. Sci.* 16 147–156 0.1016/j.tics.2012.01.005

Ryle, G. (1949/2009). *The concept of mind*. Routledge.

Robbins, P., & Jack, A. I. (2006). The phenomenal stance. *Philosophical studies*, *127*(1), 59-85.

Schilbach L., Eickhoff B., Rotarska-Jagiela A., Fink G. R., Vogeley K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Conscious. Cogn.* 17457–467 0.1016/j.concog.2008.03.013

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127-137.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., ... & Vohs, K. D. (2014). Free Will and Punishment A Mechanistic View of Human Nature Reduces Retribution. *Psychological science*, 0956797614534693.

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

Small, D. A., Loewenstein, G., & Slovic, P. (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.

Sridharan, D., Levitin, D. J., & Menon, V. (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.,. 'Freedom And Resentment'. *Free Will And Reactive Attitudes: Perspectives On PF Strawson's "Freedom And Resentment"*. M McKenna and M.P. Russell. 1st ed. Burlington, VT: Ashgate Publishing, Ltd, 2012. 19-37.

Sytsma, J. (2013). The robots of the dawn of experimental philosophy of mind.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, *151*(2), 299-327.

Sytsma, J., & Machery, E. (unpublished) The relevance of folk intuitions to evaluating the justification for the 'hard problem"

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and phenomenological research*, *76*(1), 138-155.

Uddin, L. Q., Clare Kelly, A. M., Biswal, B. B., Xavier Castellanos, F., & Milham, M. P. (2009). Functional connectivity of default mode network components: correlation, anticorrelation, and causality. *Human brain mapping*,*30*(2), 625-637.

Van Overwalle F. (2009). Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp.* 30 829–858 10.1002/hbm.20547

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

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*(3), 564-584.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2008). Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *Journal of neurophysiology*, *100*(6), 3328-3342.

Vincent, J. L., Patel, G. H., Fox, M. D., Snyder, A. Z., Baker, J. T., Van Essen, D. C., ... & Raichle, M. E. (2007). Intrinsic functional architecture in the anaesthetized monkey brain. *Nature*, *447*(7140), 83-86.

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

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

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813-836.

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