This is an Accepted Manuscript of an article published by Taylor & Francis in Neurocase on 2011, available online: http://www.tandfonline.com/https://doi.org/10.1080/13554794.2010.532808 .

"The self" from philosophy to cognitive neuroscience

Winston Chiong, MD PhD, University of California, San Francisco School of Medicine

ABSTRACT: Neuroscientists have recently begun to explore topics, such as the nature of the self, that were previously considered problems for philosophy rather than for science. This article aims to provide a starting point for interdisciplinary exchange by reviewing three philosophical debates about the nature of the self in light of contemporary work in cognitive neuroscience. Continental rationalist and British empiricist approaches to the unity of the self are discussed in relation to earlier work on split-brain patients, and to more recent work on "mental time travel" and the default mode network; the phenomenological movement, and the central concept of intentionality, are discussed in relation to interoceptive accounts of emotion and to the mirror neuron system; and ongoing philosophical debates about agency and autonomy are discussed in relation to recent work on action awareness and on insight in clinical populations such as addicts and patients with frontotemporal dementia.

KEYWORDS: Philosophy; Self; Autonomy; Personal; Consciousness; Existentialism; Memory; Default mode; Mental time travel; Mirror neuron; Intentionality

As the articles in this issue of *Neurocase* illustrate, cognitive neuroscientists have begun to address topics traditionally regarded as questions for philosophy rather than for science; such topics include not only the nature of the self, but also the relationship between the material brain and conscious mind (Block, 2005), the problem of free will (Roskies, 2010), and the nature of moral motivation (Haidt, 2007). There are many reasons for scientists interested in these matters to examine the long history of philosophical inquiry into the self when beginning their empirical

investigations. In many cases we may find that earlier philosophers have provided the clearest delineation of paradoxes and problems for further inquiry, even when techniques to answer them were not yet available. In other cases, attention to prior philosophical debates may provide a map of intellectual space. For example, if two independently compelling claims are logically inconsistent with one another, then models that attempt to accommodate both claims must be dismissed as incoherent. A third, "therapeutic" role for attention to the history of philosophy was suggested by Wittgenstein and has been famously adopted by Daniel Dennett (1991): we may discover that outdated philosophical models continue to underlie our unexamined assumptions about some phenomenon. Alternatively, even discredited or incomplete philosophical models may serve as sources of inspiration that point towards innovative ways of thinking about familiar problems.

At the same time, there are also dangers in interdisciplinarity, particularly if scientists attempt to straightforwardly apply philosophers' positions to their own areas of interest without regard for their original contexts. One such pitfall is an anachronistic one: failing to note when the background assumptions of earlier theorists do not match our own. While contemporary cognitive neuroscientists generally accept that adequate explanations of psychological phenomena must ultimately cohere within a general account of a world governed by microphysical laws, earlier philosophers would have been less constrained by this scientific worldview. At the same time, we are more comfortable than most ancients and early moderns with pure contingencies; that is, with the idea that our explanations may simply come to an end with a statement of fact that itself has no deeper or unifying explanation. For instance, we are more willing to entertain the possibility that consciousness does not serve a biological or theological function, and more fundamentally, we no longer expect human interests (or a teleology interpretable by analogy to human interests) to figure

in the most basic levels of explanation. There is a further problem that arises in attempting to bridge philosophical investigation of the self and contemporary cognitive neuroscience. Some areas of philosophical inquiry are well-demarcated, with general agreement in terminology and in the overall boundaries of debate (for instance, philosophical debates over knowledge, justice, how terms secure reference, or the ontological status of biological species). However, there is no well-demarcated "philosophy of the self" – philosophers use this term in many different ways, and in order to address many different questions. There are, then, many overlapping philosophical debates that pertain to commonsense notions of identity, personhood, reflexivity, and individuality. Whether there is any unified account of the self that can link these many different (but presumably related) topics is itself a matter for philosophical and neuroscientific inquiry.

In the remainder of this essay, I will attempt to summarize three philosophical debates pertaining to the self - the first on rationalist and empiricist accounts of the perceiving subject, the second on phenomenology, and the third on the nature of intentional action. I will focus on their relevance, respectively, to the unity of the self, the relationship between the self and the external world in perception, and the relationship between the self and the external world in action; and then in each case will turn to consider contemporary neuroscientific work that intersects with these philosophical discussions. Given the breadth of the subject, this review does not aim at a comprehensive survey of philosophical and neuroscientific approaches to the self, but should serve to illustrate the variety of philosophical topics that pertain to our commonsense notion of the self, and to provide an entry point for deeper study of some of the issues summarized here (Figure 1).

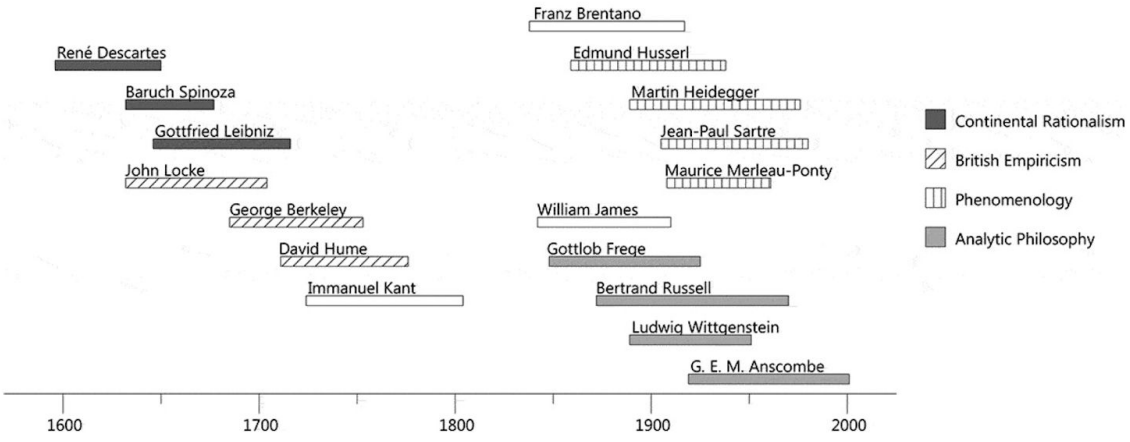

Figure 1: Historical philosophers discussed in text, and related thinkers, organized by lifetime and (roughly) by philosophical tradition.

Rationalism, Empiricism, and the Unity of the Perceiving Subject

Continental rationalism and British empiricism, as positions in the history of philosophy, are principally concerned with the sources of knowledge and epistemic justification; with empiricists insisting upon sense experience as the ultimate source of knowledge, and rationalists instead claiming that reason alone can provide us with knowledge that is is *a priori,* or independent of experience. These epistemological disputes are beyond the scope of this article; however, it is useful to note how these philosophers' arguments about reason and experience are shaped by their influential views about human minds, which are after all both the bearers of rational faculties and the subjects of sense experience.

René Descartes (1641) begins the "epistemological turn" in modern philosophy by seeking an absolutely secure foundation for all knowledge. He rejects sense experience as a reliable guide to truth, citing visual illusions, dreams, hallucinations and phantom limb pains as examples in which

sense experiences deceive us about the nature of the world. However, he claims that even if he were deceived in all of his thoughts and experiences, *he* must still exist as the subject of these thoughts and experiences – "I think, therefore I am". Notably, Descartes here can only express certainty in his existence as a thinking being, whereas his physical attributes as reported by his senses remain in doubt. This anticipates his dualistic separation of mind and body as separate substances.

While philosophers and neuroscientists have since tried to avoid Descartes's substance dualism, one of Descartes's ancillary arguments for substance dualism has continued implications for our sense of self. Descartes notes that it is in the nature of physical bodies, which are extended in space, to be divisible; however, he claims that the mind considered as a thinking substance is indivisible, and therefore cannot be a physical body. Descartes's actual argument is difficult to unpack, as it takes up only a paragraph; however, if we focus purely on the mind as the bearer of conscious states, there is something powerfully intuitive about this claim. My conscious experience feels to me to be unified, such that all of my individual conscious experiences are equally and immediately available to a single subject (namely, myself), and it is difficult to conceive of how these experiences could be realized in distinct and separable parts of my brain. While split-brain studies demonstrate that consciousness can in fact be divided even within a single human being (Gazzaniga, 2000), Thomas Nagel (1971) argues that we have yet to develop a theoretical framework and self-conception that can help us to fully make sense of these findings. In this way, Cartesian intuitions about the unity of consciousness continue to have a powerful hold on our thinking about the mind.

While Descartes seeks to base our knowledge in reason, John Locke attempts to show that experience, rather than reason, is the source of our concepts and knowledge (with some exceptions,

such as our knowledge of God's existence) (Locke, 1690). Locke begins with a conception of the mind at birth as a blank slate, and attempts to show how sense impressions provide the mind with simple ideas that can then be built up into more complex ones. Locke's view, like Descartes's, requires the existence of a perceiving subject that persists over time; but unlike Descartes, Locke does not understand this conscious subject as a separate substance distinct from its sense experiences. For Locke, what ties together my current conscious experiences with (for instance) my conscious experiences from a year ago is not that they are both possessed by the same immaterial soul; instead, they are linked by the relations of memory that hold between these mental states.

For Locke, then, the self is constituted by memory, and what makes me the same person as I was a year ago is the fact that I now can recall my experiences from a year ago. Locke argues that, even if our consciousness were the product of a separable immaterial substance, my persistence as the same person over time would still depend on memory rather than on the persistence of a Cartesian soul. For instance, if we accept reincarnation, then the same immaterial soul might possess conscious experiences over the course of many different human lives; yet if the bonds of memory are broken between one life and the next, then these experiences would not belong to the same *self* even if they are experienced by the same soul. However, if a soul were to move from one body to another while keeping the memories of its prior life, Locke claims that the resulting person would be identical with the prior person who had these past experiences.

An obvious problem for Locke's view is that it implies that, if I forget some experience, then that experience was not mine (Hume, 1739; Reid, 1785). While Locke attempts to build a conception of the self and personal identity on empiricist grounds, David Hume claims to find no basis in experience for our belief in the existence of a persisting subject. As he writes,

> when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a perception, and never can observe anything but the perception. (Hume, 1739, p. 252)

On Hume's deflationary "bundle theory" of the self, there is nothing more to the perceiving subject than a bundle of different sense experiences, "in a perpetual flux and movement," and without any underlying unity. However, Hume found himself unable to explain how these bundles of experience are brought together such that, for instance, all of *my* disparate experiences constitute one bundle while all of *your* experiences constitute a different bundle. In an Appendix written 18 months later, Hume identifies this as the one area in his philosophy in which he has found considerable mistakes, and with atypical modesty he ultimately gives up the problem of personal identity as too difficult for his understanding.

Contemporary philosophers have returned to the core of Locke's account of personal identity, while amending it in response to earlier objections. Derek Parfit (1984) and Sydney Shoemaker (1984), among others, present more complicated variations of Locke's view that allow a richer conception of the different relationships that can hold between conscious experiences at different times even in the absence of direct memory connections. Such relationships need not be solely retrospective, as in the case of memory, but can also be prospective, as in the case of intention. Intentions also demonstrate how prospective and retrospective mental states may depend on one another; among other things, my intention now to go for a walk this evening will only be effective if, later this evening, I remember and carry out this prior intention. (For instance, if I forget about this intention but later go out for a different reason, this would not fulfill my prior intention.) On neo-Lockean views of personal identity, mental states like memory and intention that link temporally discontinuous experiences do not merely allow awareness of my self over time; instead,

they *constitute* my self, in part by grounding the capacity for temporally-extended agency (Bratman 1987, 2000).

Recent work in cognitive neuroscience has uncovered intriguing links between retrospective and prospective mental states, suggesting how they may function together to constitute a stable subject of experience over time. For instance, while it has long been recognized that patients with bilateral hippocampal injury have profound impairments in episodic memory, more recent studies demonstrate that such patients have similar impairments in imagining new experiences (Hassabis, Kumaran, Vann, & Maguire, 2007). In healthy subjects, the phenomenological richness of both recalled past and imagined future events decreases with temporal distance from the present (D'Argembeau & Van der Linden, 2004); functional neuroimaging studies demonstrate overlapping patterns of brain activation for recollecting the past and envisioning the future (Okuda et al., 2003; Szpunar, Watson, & McDermott, 2007); and child development studies indicate that these capacities emerge in tandem (Perner, Kloo, & Rohwer, 2010). These findings suggest that episodic recall and prospective imagery both depend on a common faculty of "mental time travel" (Suddendorf & Corballis, 1997; Tulving, 1983), subserved by a network with prominent components in the medial prefrontal cortex, precuneus, posterior cingulate cortex, and medial temporal lobes, which allows us to re-experience past events and to prospectively imagine future events. This ability to project ourselves backwards into previously encountered situations and forwards into novel situations is considered by some authors to be a distinctively human capacity (Suddendorf & Corballis, 1997) essential for social cooperation and other behavior directed beyond immediate present rewards (Boyer, 2008).

Patterns of brain activation associated with mental time travel also demonstrate remarkable overlap with other cognitive domains that are intuitively associated with the self, such as decision-making in personal moral dilemmas (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) and theory of mind (Gallagher & Frith, 2003). Yet the most intriguing functional association of retrospection and prospection may be with the default mode network (Buckner & Carroll, 2007; Raichle et al., 2001). This network was initially characterized somewhat serendipitously through functional neuroimaging data collected from subjects in a state of undirected wakefulness, often acquired by investigators as a "control" or "rest" condition against which various task-directed states could be compared. Subsequent analyses of this supposed "rest" condition revealed the intrinsic activity of a highly coherent constellation of brain regions, with metabolic demands that dwarf those evoked by particular tasks (Raichle, 2009), and which develops in coherence during the course of human maturation (Fair et al., 2008). This network also overlaps substantially with patterns of brain activation associated with memory and prospection (Spreng, Mar, & Kim, 2009), and is preferentially disrupted in Alzheimer's disease, a neurodegenerative condition marked by early deficits in episodic memory (Greicius, Srivastava, Reiss, & Menon, 2004; Seeley, Crawford, Zhou, Miller, & Greicius, 2009). Given that the default mode network is highly active in undirected wakefulness and suppressed under many task (i.e., experimenter-directed) conditions, it is believed to be involved in self-directed cognition and mind-wandering; a recent study combining experience sampling and functional neuroimaging techniques reports that thoughts of one's personal past and future are a major focus of undirected wakefulness, and that participants' tendencies to engage in such thoughts are correlated with the functional connectivity of the medial temporal lobe with nodes of the default mode network (Andrews-Hanna, Reidler, Huang, & Buckner, 2010).

Phenomenology as the Study of Experience Itself

While the rationalists and empiricists are concerned with whether experience can provide knowledge about the external world, the later field of phenomenology is concerned with a systematic investigation of experience itself. This term reflects Immanuel Kant's distinction between "phenomena", which are objects and events as they appear in our experience, and "noumena" or things-in themselves, objects and events considered independent of the forms imposed on them by our cognitive faculties (and as such, according to Kant, unknowable to us). The phenomenological movement was deeply influenced by the introspective psychological work of Franz Brentano and William James. Brentano distinguishes "genetic psychology", which studies psychology from a third person point of view through empirical experiments and other characteristically scientific methods, and "descriptive psychology", which attempts to describe the first-person point of view (Brentano, 1890). Similarly, while James relies upon his considerable background in neural anatomy and physiology (reviewed extensively in the initial chapters of *The Principles of Psychology*), he adopts the method of introspective observation, understood as "looking into our own minds and reporting what we there discover" (James, 1890, p. 185).

One feature that distinguishes the phenomenologists from these earlier introspective psychologists is that they do not view their form of introspective study as merely one autonomous branch of inquiry, but instead regard phenomenology as the foundation of all philosophy. Edmund Husserl, the founder of phenomenology, claims that such a "pure phenomenology" goes beyond the individual subjective conscious states of particular minds at particular times, and instead analyzes the structure of experience and meaning that is common among conscious subjects. While mere

psychology studies how people happen to think, phenomenology attempts to uncover a kind of objective logic of thought.

Husserl's method is to "bracket" the question of whether our experiences correspond to any external reality, focusing our reflection on the character of conscious experience itself. The first observation is that the essential feature of consciousness is *intentionality* – conscious acts such as perceiving, judging, valuing, wishing, acting, and loving are "about" or "directed at" some object (Husserl, 1913). (This important term is unfortunately misleading in English because it has a shared Latin derivation with the common English words 'intention' and 'intentional', yet has no privileged conceptual tie with them – while intentions do in fact exhibit intentionality, the "directedness" or "aboutness" of intentionality is not restricted to intentions but is a general feature of conscious states.) Such states possess intentional objects even when they are nonveridical or even impossible: a visual hallucination of a pink spotted elephant is about a pink spotted elephant, the act of baking a chocolate cake is directed at a cake that does not yet exist, and searching for the largest prime number has as its object the largest prime number.

Importantly, then, for Husserl our experiences are not principally experiences of sense-data. I do not experience a set of tightly interwoven patches of colors in front of me and then infer an image of my yellow coffee cup; and when I walk around the table, the changing lighting and perspectival features of the cup are not presented as distinct experiences that I intellectually build into a representation of a stable object viewed from different angles. Instead, while these sense-data are components of my experience, the object of my visual experience is a yellow coffee cup. This representation is not limited to my occurrent visual impressions – for instance, I perceive the cup as having an opposite side that is out of view, and so my experience of the cup includes a "horizon" of

implicit possibilities such as an expectation of the sort of visual impression I will have if I turn it around. For Husserl the intentionality of consciousness demonstrates that our experiences are fundamentally meaningful, in that we experience things *as* things rather than as disparate subjective sense impressions, in ways that are not purely private and subjective (as in my particular visual impression of the cup viewed from one angle) but instead can be shared among conscious subjects.

Husserl's greatest successors, often grouped together as "existential phenomenologists," depart from his conception of phenomenology in various ways. One departure is their insistence that the explicit intentionality of conscious experience only takes place and has its meaning within a framework of implicit forms of engagement with objects in the world. For Martin Heidegger (1927), a more basic form of directedness is exemplified by the skilled and nondeliberate use of tools—when we are expert and absorbed in a task, the features of our tools need not be explicitly represented and in fact may seem to disappear, as we simply and prereflectively treat these tools as available for the task at hand. Maurice Merleau-Ponty (1945) draws upon case studies of phantom limbs and cortical brain injuries in developing an account of what he regards as the most basic form of intentionality. This "motor intentionality" is based in our own bodily movement yet is directed towards external objects in terms of dispositions to move in various ways such as to approach, or hold, or withdraw from, or explore them.

For related reasons, these theorists reject Husserl's method of "bracketing," with its Cartesian division between conscious experience and an external world that experience may or may not faithfully represent. With this rejection, Heidegger and Merleau-Ponty move away from conscious experience as the main subject of phenomenology, and instead study a broader phenomenon that they term "being-in-the-world," which encompasses implicit and prereflective

forms of engagement. By understanding the self as practically extended into the world, they aim to dissolve traditional philosophical dualities such as the division between mind and body, or the internal and the external; whether they succeed in this aim remains controversial. A third departure, which is tied to existentialism as a literary and aesthetic movement, is their concern with the meaning of our own existence. Whereas for Husserl, we encounter intentional objects as meaningful in our conscious experience, Heidegger claims that the meaning of our own existence is not given – instead, we must take a stand on what we essentially are.

In contemporary cognitive science, the concept of intentionality has been central to philosophical criticisms of the James–Lange theory of emotion, which was independently proposed by William James and Carl Lange in the 1880s, and which has since been adopted and modified by contemporary cognitive neuroscientists such as Antonio Damasio (in his "somatic marker hypothesis") and Bud Craig (Craig, 2002; Damasio, 1993). This theory identifies emotions with interoceptive perceptions of bodily changes; while such changes may be brought about by an external stimulus, the stimulus and its mental representation are not conceived as part of the emotion itself. As James writes,

> Our natural way of thinking about these standard emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My thesis on the contrary is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion ... we feel sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful, as the case may be. (James, 1884, pp. 189–190)

The standard philosophical objection to this theory is that it cannot satisfactorily account for the intentionality of emotions, which are typically about or directed at objects in the world – for instance, someone's grief at the death of a relative, or their fear of spiders, or their love for their

children (Pitcher, 1965). The interoceptive perceptions invoked by the James–Lange theory, however, are directed at internal bodily states rather than objects in the world, and so would appear to have the wrong intentional objects. In more recent work, Jesse Prinz (2004) attempts to answer this objection by appealing to causal theories of intentional content, according to which emotions merely *register* bodily changes as a way of representing conditions in the world.

The later existential phenomenologists have been a rich source of inspiration for contemporary cognitive models. Hubert Dreyfus (1972) appeals to Heideggerian considerations in his influential critique of earlier projects in cognitive science that analogize the brain to a digital computer; while Walter Freeman (1999) presents an account of perception and learning with substantial correspondences to Merleau-Ponty's conception of motor intentionality (Dreyfus, 2000). More recently, motor intentionality has been an influential construct in models of mirror neurons, a population of neurons originally characterized in the macaque monkey, and with an analogous "mirror system" in humans (Rizzolatti & Sinigaglia, 2007; Gallese & Sinigaglia, 2010). These neurons are active not only when a monkey itself performs a 196 CHIONG given motor act, but also when it observes another individual performing a similar act. This system utilizes an individual's own motor programs to arrive at a rapid and preconceptual interpretation of others' actions and intentions. In these models, as for Merleau-Ponty, features of the external world are represented by reference to the individual's own motor dispositions, thereby linking basic levels of perceptual processing with action.

Activity, Action, Agency, Autonomy

Turning now from the self as the subject of experience to the self as the agent of intentional action, we may begin with Ludwig Wittgenstein's famous question:

> Let us not forget this: when 'I raise my arm', my arm goes up. And the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm? (Wittgenstein, 1958, §621)

Wittgenstein calls our attention to the difference between what someone does, and what merely happens to someone (as when, for instance, a patient's arm goes up during a seizure). One theme of recent philosophical work in this area is that, at least in typical or "full-blooded" cases of human agency, there are actually several levels of action that should be distinguished.

Here again, different thinkers demarcate these levels differently, but some examples may be illustrative. Perhaps the broadest level encompasses mere *activity*, things that a person does (and not merely things that happen to her) but without an explicit goal and often unconsciously, as in aimless fidgeting or drumming one's fingers against a desk. Next, there is a basic form of purposeful *action* that is exhibited by animals, as when a spider crawls across a windowsill or a dog runs to the door at the sound of his owner's keys, and which we can explain in terms of the animal's aims. A fuller form of intentional action is exhibited in the typical actions of human agents: one important feature emphasized by Elizabeth Anscombe (1957) is that in such actions I seem to know, in an especially immediate and non-observational way, what I am doing and why I am doing it. Finally, even in the case of intentional actions, we may still ask whether these actions are performed *autonomously*. Plato provides the famous example of Leontius, who was unable to overcome a morbid desire to stare at the corpses of the executed, and was disgusted with himself as a result; a more familiar contemporary example is that of an unwilling addict, who acts intentionally but against his will in succumbing to his addiction.

I wish to focus on the last two levels as most distinctive of human agency, and as most intimately involved with our sense of self. As noted by Anscombe, one interesting feature of intentional action as exhibited by agents is that it is closely connected with self-knowledge: at least in the typical case when I raise my arm, I know that I am raising my arm and I know why I am raising my arm. Is this knowledge merely an inference derived from observing my mental states and bodily movements? Many philosophers have argued on the contrary that intention entails belief, so that if I intend to raise my arm, then I believe that I will raise my arm. Perhaps the most ambitious such account has been presented by David Velleman (1989), who argues that intentions simply are a species of self-fulfilling beliefs about what we will do. On Velleman's view, I can raise my arm by forming the spontaneous belief that I will raise my arm – 'spontaneous' because it is not based upon the sort of prior evidence that I would need to conclude that you will raise your arm. This belief, in conjunction with a desire to understand myself (that is, to behave in ways that are intelligible to me) leads me to raise my arm in accordance with this belief, and thereby make this belief true.

Other philosophers have challenged the idea that there is an internal connection between intention and self-knowledge. There appear to be cases, after all, in which I intend to so something but don't believe that I will do it (for instance, intending to win an Olympic gold medal); often these cases concern aims that are not entirely within the agent's power to achieve. Velleman counters that these cases are better described as cases of intending to *try* to do something, rather than as cases of intending to do something. But this response remains problematic, as it implies that I fulfill my intention even if I try and I fail.

Even in unambiguous cases of intentional action, these actions may or may not be undertaken autonomously. In an influential paper, Harry Frankfurt (1971) suggests that freedom of

the will, as opposed to mere free action, depends upon our reflective stance towards our motivations. He illustrates this claim with the example of an unwilling addict, who struggles against but ultimately succumbs to his desire to take some drug. While the desire for the drug is in a straightforward sense his own, there is another respect on which this desire is alien to him—and for this reason, even though the addict acts freely (in that no other person forces him to take the drug), he does not act of his own, there is another respect on which this desire is alien to him – and for this reason, even though the addict acts freely (in that no one else forces him to take the drug), he does not act of his own free will insofar as he acts on a desire that he does not want to act upon. Frankfurt contrasts this case with what he calls a wanton drug addict, who acts from his desires without reflecting upon them. In the case of wantons, Frankfurt claims, the question of whether they act of their own free will does not arise.

Unlike the unwilling addict, in most cases of human action we are not alienated from our desires, but instead identify with them. These experiences of identification and alienation indicate that, even when some motivation is unquestionably a feature of one's psychology, that motivation may nonetheless be external to the self. Frankfurt initially suggests that identification and alienation can be understood simply in terms of having second-order desires about the effectiveness of our first-order desires, but this view has proven problematic. In addition to the problem of regress, it is not clear why one desire should enjoy priority over another desire in representing the perspective of the self, simply in virtue of being higher-order (Watson, 1975). While many philosophers share Frankfurt's core intuition that identification and alienation involve some form of reflection upon one's own motives, the link between self-reflection and this form of autonomy remains opaque.

Returning to Anscombe's claim that agents have non-observational knowledge of their own actions, recent neuroscientific work on action awareness has generated intriguing findings about the links between intention, bodily movement, and awareness of movement. Desmurget et al. (2009) report a curious dissociation of these three conditions in a study of patients undergoing awake electrical stimulation of the brain during brain tumor surgery (a functional mapping technique used to minimize the risk of postoperative neurological deficits). Patients stimulated in posterior parietal cortex (Brodmann areas 39 and 40) at low intensities reported a felt desire to move (interpreted by the authors as an "intention"); when these patients were again stimulated at higher intensities, they reported the false belief that they had moved, in the absence of overt movement or EMG activity. Meanwhile, patients stimulated in premotor cortex (dorsal Brodmann area 6) exhibited complex multijoint movements but denied that they had moved. This contrasts with an earlier study in which stimulation of the supplementary motor area (mesial Brodmann area 6) at low intensities elicited a subjective "urge" to move or anticipation that a movement would occur, and where stimulation at higher intensities resulted in an overt movement (Fried et al., 1991). Desmurget and colleagues' findings suggest that, in ordinary cases, the belief that one is carrying out an action is not based upon proprioception or other self-observation, since their subjects exhibit belief in the absence of movement and movement in the absence of belief. To account for these findings, Desmurget and Sirigu (2009) propose a parietal-premotor network for intentional action, in which movement intentions are generated in the posterior parietal cortex and project to the supplementary motor area (and then to primary motor cortex) in order to generate movement; while in parallel a predictive forward model is generated in the posterior parietal cortex that is the neural correlate of movement awareness and that does not rely upon somatosensory feedback unless expectations generated in

premotor cortex are grossly violated. On this model, at least in normal cases of simple bodily movement, there is indeed an internal link between intention and self-knowledge.

Meanwhile, Frankfurt's suggestion that self-reflection is central to the will might be elucidated by studying clinical populations in which disorders of the will are associated with deficiencies of insight and self-awareness. As in Frankfurt's examples, one such population is patients with drug and alcohol addiction, in whom lack of insight into the severity or consequences of addiction is believed to play an exacerbating role (Goldstein et al., 2009). Moeller and colleagues (2010) report that cocaine addicts have impaired insight relative to controls in characterizing their own response patterns in a choice task involving drug-related and drug-unrelated imagery; and that, in those patients (a majority) without cocaine metabolites in their urine at the time of testing, insight was inversely associated with drug spending. Another such population is patients with behavioral variant frontotemporal dementia, a neurodegenerative condition marked by behavioral changes such as violations of social and ethical norms, decline in personal hygiene and grooming, binge eating, hoarding, and impulsivity. In many ways, these behaviors suggest the unreflective appetitive behavior of Frankfurt's wantons. It is noteworthy, then, that loss of insight is one of the core diagnostic features of this variant of frontotemporal dementia (Neary et al., 1998), as confirmed in studies of self-awareness for personality change (Rankin, Baldwin, Pace-Savitsky, Kramer, & Miller, 2005) and self-conscious emotion (Sturm, Rosen, Allison, Miller, & Levenson, 2006) in these patients. Further behavioral and neuroanatomical characterization of these patients' deficits may offer insights into a normal human motivational structure that is central to our sense of self.

References

Andrews-Hanna, J. R., Reidler, J. S., Huang. C., & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology, 104,* 322–335.

Anscombe, G. E. M. (1957). *Intention*. London, UK: Basil Blackwell.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences, 9,* 46–52.

Boyer, P. (2008). Evolutionary economics of mental time travel? *Trends in Cognitive Sciences, 12,* 219–224

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (2000). Reflection, planning, and temporally extended agency. *The Philosophical Review*, *109*, 35-61

Brentano, F. C. (1890). *Descriptive psychology*. Translated and edited by B. Müller. Oxford, UK: Routledge, 1995.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11,* 49–57.

Churchland, P. S. (1986) *Neurophilosophy: Towards a unified science of the mind-brain.* Cambridge, MA: MIT Press.

Craig, A. B. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience, 3*, 655–666.

Damasio, A. R. (1993). *Descartes' error: Emotion, reason and the human brain.* New York, NY: Putnam.

D'Argembeau, A., & Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition, 13*, 844–858.

Dennett, D. (1991). *Consciousness explained.* Boston, MA: Little, Brown & Company.

Descartes, R. (1641). *Meditations on first philosophy*. Translated by D. A. Cress. Indianapolis, IN: Hackett Publishing Company, 1980.

Desmurget, M., & Sirigu, A. (2009). A parietal-premotor network for movement intention and motor awareness. *Trends in Cognitive Sciences, 13,* 411–419.

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement intention after parietal cortex stimulation in humans. *Science, 324,* 811–813.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. New York, NY: MIT Press.

Dreyfus, H. L. (2000). A Merleau-Pontyian critique of Husserl's and Searle's representationalist accounts of action. *Proceedings of the Aristotelian Society, 100*, 287–302.

Fair, D. A., Cohen, A. L., Dosenbach, N. U., Church, J. A., Miezin, F. M., Barch, D. M., et al. (2008). The maturing architecture of the brain's default network. *Proceedings of the National Academy of Sciences, 105*, 4028–4032.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. Reprinted in *The importance of what we care about* (pp. 11-25). Cambridge, UK: Cambridge University Press, 1988.

Freeman, W. J. (1999). *How brains make up their minds*. London, UK: Weidenfeld & Nicolson.

Fried, I., Katz, A., McCarthy, G., Sass, K. J., Williamson, P., Spencer, S. S., & Spencer, D. D. (1991). Functional organization of human supplementary motor cortex studied by electrical stimulation. *The Journal of Neuroscience, 11*, 3656–3666.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7*, 77–83.

Gallese, V., & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia, 48*, 746–755.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain, 123,* 1293–326.

Goldstein, R. Z., Craig, A. D., Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., et al. (2009). The neurocircuitry of impaired insight in drug addiction. *Trends in Cognitive Sciences, 13,* 372–380.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Greicius, M. D., Srivastava, G., Reiss, A. L., Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proceedings of the National Academy of Sciences, 101*, 4637–4642.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316,* 998–1002.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences, 104*, 1726–1731.

Heidegger, M. (1927) *Being and time*. Translated by J. Macquarrie & E. Robinson. New York, NY: Harper & Row, 1962.

Hume, D. (1739). *A treatise of human nature*. Edited by L. A. Selby-Bigge, revised by P. H. Nidditch. Oxford, UK: Clarendon Press, 1978

Husserl, E. (1913) *Ideas: General introduction to pure phenomenology*. Translated by W. R. Boyce Gibson. London, UK: George Allen & Unwin Ltd., 1931.

James, W. (1884). What is an emotion? *Mind, 9,* 188–205.

James, W. (1890). *The principles of psychology*, Volume 1. New York, NY: Henry Holt & Company.

Lewis, D. K. (1983). *Philosophical papers: Volume I*. Oxford, UK: Oxford University Press.

Locke, J. (1690). *An essay concerning human understanding*. Abridged and edited by A. D. Woozley. London, UK: W. Collins, 1964.

Merleau-Ponty, M. (1945) *Phenomenology of perception*. Translated by C. Smith. London, UK: Routledge & Kegan Paul, 1962.

Nagel, T. (1971). Brain bisection and the unity of consciousness. Reprinted in *Mortal Questions* (pp. 147-164). Cambridge, UK: Cambridge University Press, 1979.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., et al. (1998). Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria. *Neurology, 51*, 1546–1554.

Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Tanji., K., Suzuki, K., et al. (2003). Thinking of the future and past: The roles of the frontal pole and the medial temporal lobes. *Neuroimage, 19*, 1369–1380.

Parfit, D. (1984). *Reasons and persons*. Oxford, UK: Oxford University Press.

Perner, J., Kloo, D., & Rohwer, M. (2010). Retro-and prospection for mental time travel: Emergence of episodic remembering and mental rotation in 5- to 8-year old children. *Consciousness and Cognition*. Advance online publication. doi:10.1016/j.concog.2010.06.022.

Pitcher, G. (1965). Emotion. *Mind, 74*, 324–346.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford, UK: Oxford University Press.

Raichle, M. E. (2009). A paradigm shift in functional brain imaging. *The Journal of Neuroscience*, 29, 12729–12734.

Raichle, M. E., MacLeod, A. M, Snyder, A. Z., Powers, W. J., Gusnard, D.A., & Shulman G. L. (2001) A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*, 676–682.

Rankin, K. P., Baldwin, E., Pace-Savitsky, C., Kramer, J. H., & Miller, B. L. (2005). Self awareness and personality change in dementia. *Journal of Neurology, Neurosurgery & Psychiatry, 76*, 632–639.
Rawls, J. (1971). A theory of justice. Cambridge, MA: Harvard University Press.

Reid, T. (1785). Of Mr. Locke's account of our personal identity. Reprinted in J. Perry (Ed.), *Personal identity* (pp. 113-118). Berkeley, CA: University of California Press, 1975.

Rizzolatti, G., & Sinigaglia, C. (2007). Mirror neurons and motor intentionality. *Functional Neurology*, *22*, 205–210.

Roskies, A. L. (2010). How does neuroscience affect our conception of volition? *The Annual Review of Neuroscience*, 33, 109–130.

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., & Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron, 62*, 42–52.

Shoemaker, S. (1984). Personal identity: a materialist's account. In S. Shoemaker & R. Swinburne (Eds.), *Personal identity* (pp. 67-132). Oxford, UK: Blackwell Publishers.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*, 489–510.

Sturm, V. E, Rosen, H. J., Allison, S., Miller, B. L., & Levenson, R. W. (2006). Self-conscious emotion deficits in frontotemporal lobar degeneration. *Brain, 129*, 2508–2516.

Szpunar, K. K., Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences, 104*, 642–647.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs*, *123*, 133-167.

Tulving, E. (1983). *Elements of episodic memory*. Oxford, UK: Clarendon Press.

Velleman, J. D. (1989). *Practical reflection*. Princeton, NJ: Princeton University Press.

Watson, G. (1975). Free agency. Reprinted in G. Watson (Ed.), *Free will* (pp. 96-110). Oxford, UK: Oxford University Press, 1982.

Wittgenstein, L. (1958). *Philosophical Investigations* (3rd English ed.). Translated by G. E. M. Anscombe. New York, NY: Macmillan Publishing Co.