CHAPTER 5

Self-Knowledge

From Philosophy to Neuroscience to Psychology

MATTHEW D. LIEBERMAN

The contemplation of self-knowledge has a long history within philosophy, a comparatively short history within psychology, and a vanishingly short history within neuroscience. Thus, if we wish to see whether neuroscience has something useful to tell us about the self, we would do well to situate it in the context of the 2,000-year dialogue that philosophers have had about the nature of the self and the paths by which we can know something about it.

This chapter is divided into two parts. In the first part, I consider various pronouncements by philosophers about the nature of self-knowledge and see what neuroscientists have had to say that supports or conflicts with those claims. In the second part, I start from the brain region, medial prefrontal cortex (MPFC), that has mostly commonly been associated with self-knowledge and self-related processes more generally, and examine the other psychological processes that invoke this region in order to try to make some progress in identifying the kinds of processes that are invoked when we think about the self or retrieve self-knowledge.

To avoid turning this into a mystery novel, here's the punchline. First, the MPFC may be more commonly activated by self-processing than anything else that has been studied; however, self-processing may just be a very prominent example of a broader class of processes. This is not to suggest that knowing the self is just like knowing *anything* else we know well (Greenwald & Banaji, 1989) because this is clearly not the case (Macrae, Moran, Heatherton, Banfield, & Kelley, 2004).

Vazire.indb 63 3/1/2012 11:12:47 AM

From Philosophy to Neuroscience

Ancient Greece Invokes the Challenge

The Ancient Greeks had profound influence over the course that Western civilization has taken over the past two millennia. Who would have guessed that the inscription "know thyself" above the Oracle at Delphi or Socrates' warning that the "unexamined life is not worth living" would have led to such an onslaught of interest, leading ultimately to countless self-help aisles in second-rate mall bookstores full of advice on how to understand oneself and how to use those insights to hack into one's own operating system and make changes for the better?

In 2002, Kelley and colleagues (2002) published the first neuroimaging paper on self-knowledge that identified the key region, MPFC, that has been the center of virtually every self-related functional magnetic resonance imaging (fMRI) study since (see Craik et al., 1999, for an earlier attempt). The paradigm was simple and straightforward. Participants were presented with a trait word such as *dependable* or *polite* on each of a series of trials. Additionally, on the same screen was an instruction cue that indicated the task to be performed for that trial. On some trials, participants indicated whether the trait word was self-descriptive. On other trials, participants indicated with whether the trait word described George Bush, a well-known public figure, and on still other trials, participants determined whether the trait word was presented in uppercase or lowercase letters (a task originally developed by Markus [1975] and Rogers et al. [1977]).

Two regions were more active during self-reference trials compared to those that required referencing knowledge about another person: MPFC and an overlapping region of precuneus and posterior cingulate cortex (i.e., precuneus$_{PCC}$).

In the last decade, about three dozen neuroimaging studies of self-knowledge (or self-reference, as it is typically described in the neuroimaging literature) have been conducted. In a review of 32 of these studies (Lieberman, 2010; see also van Overwalle, 2009), MPFC was observed in 94% of the studies, with precuneus$_{PCC}$ appearing in 63% of the studies. The only other region appearing in more than half of the studies, at 53%, was dorsomedial prefrontal cortex (DMPFC), a region commonly associated with thinking about the mental states of others.

If we are to heed the instruction to "know thyself" from a neuroscience perspective, this first study provided the bedrock, the foundation upon which to build. But the studies using variants of this paradigm have also led to as many questions as answers. It is also important to note a terminological difference between self-knowledge research within neuroscience and social psychology. Within social-cognitive neuroscience, encoding new information about the self (coming to know the self), making judgments about the self, or retrieving self-related information are all considered self-knowledge processes. In contrast, within social psychology, self-knowledge more frequently refers to the accuracy of people's beliefs about themselves. Neuroscientists would probably refer to this as self-insight or the accuracy of self-beliefs, rather than self-knowledge per se. Regardless of the terminology, there is precious little research on the neural bases of self-knowledge accuracy, but what little there is, is quite consistent with self-reference findings, implicating MPFC in this as well (Beer, John, Scabini, & Knight, 2006; Schmitz, Rowley, Kawahara, & Johnson, 2006; Schnyer,

Nicholls, & Verfaelli, 2005). For the remainder of this chapter, however, I use *self-knowledge* to refer more broadly to all aspects of knowledge about the self, including the processes involved in accessing this information.

Enlightenment and Temporality of the Self

In 1689, John Locke, a British empiricist and one of the intellectual giants of the early Enlightenment period, included a chapter in *An Essay Concerning Human Understanding* entitled "Identity and Diversity." Here, he wrote about memory as the basis on which identity stands, concluding that "as far as consciousness can extend backwards in time to any past action or thought, so far reaches the identity of that person" (1689/1975, p. 335). He immediately appreciated the neuropsychological implication of his claim:

> But yet possibly it will still be objected: Suppose I wholly lose the memory of some parts of my life, beyond a possibility of retrieving them, so that perhaps I shall never be conscious of them again; yet am I not the same person that did those actions, had those thoughts that I once was conscious of, though I have now forgot them? To which I answer, that we must here take notice what the word "I" is applied to; which, in this case, is the man only. And the same man being presumed to be the same person, "I" is easily here supposed to stand also for the same person. (p. 342)

He then went on to suggest that this final supposition is incorrect; while an amnesic is literally the same man, Locke contends he is no longer the same person, no longer in possession of the same identity, because he cannot recall the information that was the basis of his former identity. He was not suggesting that the man is not the same because he has changed in some minor fashion. Rather, he claimed that these are two distinct individuals because there is no overlap in their memories.

Klein, Loftus and Kihlstrom (1996) were able to examine Locke's claim directly in one of the first social cognitive neuroscience experiments ever conducted. They studied a patient who was temporarily amnesic due to head injury. They found that this patient possessed trait self-knowledge that was largely equivalent to that observed outside of the amnesic state. She made trait ratings of herself during her amnesic state and afterwards, and these two sets of ratings correlated .74, almost identical to the test–retest correlation of control subjects ($r = .78$). In other words, even though this patient could not remember the episodes in her life that would have led her to believe she was, say, generous, she still knew she was a generous person.

In separate behavioral research, Klein, Loftus, Trafton, and Fuhrman (1992) observed that remembering specific episodes of past behavior was only relevant to making self-judgments in domains where one had relatively little experience. Specifically, subjects were asked to remember trait-specific memories immediately prior to making trait self-judgments, and this only facilitated the speed of self-judgments in low-experience domains. Lieberman, Jarcho, and Satpute (2004) followed up this work with an fMRI study comparing self-knowledge (i.e., self-reference) in high- and low-experience domains. Consistent with Klein and colleagues, this study revealed that the medial temporal lobe, central to the storage of episodic memories, was only involved in self-knowledge for low-experience domains. In contrast, high-experience

domain self-knowledge was associated with MPFC and precuneus$_{PCC}$ along with other regions associated with more automatic or implicit processes. Together, these studies suggest that, on the whole, as brilliant as he was, Locke had it wrong about the relationship of memory to identity. Except perhaps in new domains of experience, our sense of self outlives the memories that may have given rise to the original self-insights (see also Hastie & Park, 1986).

John Butler, a lesser known philosopher from the Enlightenment period, who wrote the chapter "Of Personal Identity" in his book *The Analogy of Religion* (1736/1819 [reprinted in *Works*, 1896]), goes in quite the opposite direction from Locke, suggesting that memory is not enough to forge together the identities of the same man at two different points in time. Butler takes the "ever-changing river" approach to the self, suggesting that states of consciousness are what define the self, and just as a river is never the same at two different points in time, neither is consciousness, and thus neither is the self. He wrote:

> No one can any more remain one and the same person two moments together than two successive moments can be one and the same moment. . . . And from hence it must follow that it is a fallacy upon ourselves to charge our present selves with anything we did . . . or that our present self will be interested in what will befall us tomorrow. (p. 213)

On the one hand, this statement comes across as semantic hairsplitting, but from a psychological perspective there does seem to be some truth to the fact that, at times, we do treat our future and past selves as distinct from the current self (Libby & Eibach, 2002). Oftentimes an apology amounts to being distant enough from the actions of a past self that one can promise such actions will never happen again. It is as if someone else was responsible for those actions, and we promise not to bring that irresponsible person around again. On the flip side, *temporal discounting* is the study of the fact that we are much less motivated by the pleasures and pains of our future self compared to those we can receive today (Loewenstein & Elster, 1992).

fMRI research on past and future selves backs up Butler's account, at least from a psychological, if not an ontological, perspective. Simply put, MPFC is more active when one reflects on the current self rather than either the past or future self. D'Argembeau, Xue, Lu, Van der Linden, and Bechara (2008) asked individuals to indicate whether traits were self-descriptive or other-descriptive, and varied whether participants were answering about the target now or from a prior period of time. The only three regions that were more active when thinking about the current self than any of the other combinations were MPFC, DMPFC, and precuneus$_{PCC}$. A follow-up study that included past, present, and future perspectives on the self (D'Argembeau et al., 2010) observed greater MPFC activity for the present compared to the past and future perspectives (see also Ersner-Hershfield, Wimmer, & Knutson, 2009). Similarly, work on mindfulness meditation has found that prior to mindfulness training, individuals who are very caught up in the self of the moment show much greater MPFC activity than after being trained in mindfulness to take a more detached view of the self (Farb et al., 2007; see also Way, Creswell, Eisenberger, & Lieberman, 2010). In contrast, lateral parietal regions were more active when considering the self in other time periods, consistent with another recent study on imagining the self

taking a walk in past, present, or future periods (Nyberg, Kim, Habib, Levine, & Tulving, 2010).

Sources and Components of Self in Modern Philosophy

The modern period of philosophy is commonly dated to middle of the 19th century. There are countless philosophers from this period who have weighed in on the nature of the self and our knowledge of it, but unfortunately social cognitive neuroscience research has yet to yield studies relevant to most of them. I am going to cheat a little here by turning to William James, who is often considered the founder of modern experimental psychology in America but was first and foremost a philosopher. In his work, *Psychology*, an abbreviated version of his massive two-volume opus *Principles of Psychology*, James posited a key division between components of the self:

> The consciousness of Self involves a stream of thought, each part of which as "I" can remember those which went before, know the things they knew, and care paramountly for certain ones among them as "Me." . . . This Me is an empirical aggregate of things objectively known. The I which knows them cannot itself be an aggregate. (1892, p. 215)

Similarly, in *Principles* he discussed the two elements of the self as "an objective person, known by a passing subjective thought and recognized as continuing in time. Hereafter, let us use the words ME and I for the empirical person and the judging Thought" (1890, p. 371).

According to this account, an active part of the self is involved in experiencing the world, one's phenomenological point of view, and reflecting on the passive part of the self that represents our repository of self-knowledge. There is a special file cabinet of self-knowledge called the ME, and the I is what fills the file cabinet and can later peruse its contents. Given how computationally distinct these two components of the self would have to be, one might naturally expect to see different brain regions involved in each.

Very few of the fMRI studies on self-knowledge can address this question because most confound the act of self-reference (i.e., the I reflecting on the self) with activating self-knowledge (i.e., the ME that is reflected upon). A few studies have made some attempt to separate these, but they yield somewhat different conclusions. One study took a developmental approach, comparing adults and 10-year-old children as they made self-referential judgments (Pfeifer, Lieberman, & Dapretto, 2007). Here, the assumption was that to the extent the I and ME are separable, the I's act of retrieving self-knowledge might be more automatic in adults than in children but the ME, the repository of self-knowledge, would be much more developed and consolidated (see also Wang, Lee, Sigman, & Dapretto, 2006). Thus, we hypothesized that regions more active during self-reference in children would correspond more to the I, and regions that were more active in adults might correspond more to the ME. MPFC and precuneus$_{PCC}$ were two of the only regions that were more active in the children than in the adults. In contrast, the lateral temporal cortex and angular gyrus were the only regions more active in adults than in children. Similarly, Blakemore, den Ouden, Choudhury, and Frith (2007) asked people to imagine their intentions in various

situations and found that adolescents activated MPFC more than adults, whereas adults activated superior temporal sulcus to a greater degree.

From these two studies alone, one might conclude that MPFC is more associated with the act of self-reflection than with the contents of self-knowledge. This would be consistent with the view that the prefrontal cortex is generally more involved in orchestrating information and behavioral responses elsewhere in the brain than in storing that content directly. Unfortunately, the other two relevant studies suggest the opposite conclusion. Both of these studies (Moran, Heatherton, & Kelley, 2009; Rameson, Satpute, & Lieberman, 2010) compared explicit self-knowledge, during which individuals explicitly reflected on the self, to implicit self-knowledge, during which self-relevant images were presented without any instruction to consider their self-relevance. To the extent that the I and the ME, in James's formulation, are separable, one would expect I-specific activations to be absent in the implicit conditions. Instead, both studies found significant overlap in the MPFC region recruited by both explicit and implicit tasks. Rameson and colleagues (2010) also found overlap in $\text{precuneus}_{\text{PCC}}$ across the two tasks. In other words, these two studies suggest that MPFC is involved in the representation of self-knowledge, not just the manipulation of self-knowledge through self-reflection processes. Combined with the previous two studies, the only thing we can conclude is that the jury is still out and more research is needed.

Gilbert Ryle, a 20th-century British philosopher, was described as both a behaviorist and a phenomenologist, sometimes in reference to the same work. In his most famous work, *The Concept of Mind*, he devoted an entire chapter to self-knowledge, in which he concluded:

> The sorts of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same . . . . John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe. (1949, pp. 155–156)

These words led to Bem's (1972) influential self-perception approach to self-knowledge. Bem opened his chapter on self-perception theory with the following:

> Individuals come to "know" their own attitudes, emotions, and other internal states partially by inferring them from observations of their own overt behavior and/or the circumstances in which this behavior occurs. Thus, to the extent that internal cues are weak, ambiguous, or uninterpretable, the individual is functionally in the same position as an outside observer. (1972, p. 2)

I quote Bem at length here to point out an important distinction between Bem and Ryle. It is commonly assumed that Bem held the same position as Ryle, when in fact Bem was always careful to suggest that self-observation is only necessary when internal cues are insufficient to form a judgment. Ryle, on the other hand, implied that all self-knowledge is generated through external sources.

Although no neuroscience research to date has examined the generation of self-knowledge through observing one's own behavior, there is strong evidence that processing the self through internal and external sources of information rely on independent

neural networks. When people view external manifestations of themselves, whether it is images of their own faces or a video feed showing their own arm movements in real time, a network of lateral frontal and parietal regions, particularly on the right side of the brain, is recruited (Lieberman, 2007). In contrast, when people focus on internal characteristics such as their feelings, preferences, dispositions, and goals for the future, the characteristic MPFC and precuneus$_{PCC}$ activations are found.

This dissociation between internally focused and externally focused self-processes (Lieberman, 2007, 2010) has a number of implications for how we understand the self. First, it casts doubt on the unity of self-awareness and self-knowledge processes. By extension, it casts doubt on whether the "mirror test" of self-awareness (Gallup, 1970) is a test of generic self-awareness or is instead merely a test of physical self-recognition. Second, it may help explain why in the face of overwhelming evidence to the contrary, nearly all people behave as if mind–body dualism is true. The fact that the brain evolved separate systems for consideration of our own minds and bodies may render us incapable of experiencing them as one thing.

From Neuroscience to Psychology

As can be seen, neuroscience has been able to weigh in on at least a few of the claims made over the millennia by those who have opined most famously about the nature of self-knowledge. What else can we learn from neuroscience about the self? If we know the regions that tend to be involved in self-knowledge, what good does that do psychologists?

What Else Do We Know about MPFC?

Across the three dozen or so fMRI studies of self-knowledge (i.e., self-reference), MPFC is unequivocally the touchstone region. Publishing a paper on the neural bases of self-knowledge without activations in MPFC is likely to be difficult. Even though rostral anterior cingulate cortex (rACC) is adjacent to MPFC and self-knowledge tasks often produce activations overlapping the two regions, as an editor I have witnessed harsh reviews of papers that have rACC but not MPFC activations during self-reference. It must be pointed out that although self-reference tasks almost uniformly activate MPFC, this does not imply that MPFC is a "self" region or that activation there necessarily implies that self-referential processes are occurring.

I do think it is safe to say that the MPFC plays a uniquely human role in self and social cognition. MPFC is the only region of the prefrontal cortex that is verifiably larger in humans than in other primates after researchers control for brain and body size (Semendeferi et al., 2001). Moreover, this region has disproportionately greater spacing between neurons than in other primates, thought to allow for more complex connectivity (Semendeferi et al., 2011). Thus there is something distinctive about this region in humans compared with other species.

A variety of studies provide the link from MPFC to social cognition, rather than to self-processes only. For instance, Mitchell and colleagues (Mitchell, Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006) have found in a number of studies that when individuals judge the psychological characteristics of those similar to

themselves they recruit MPFC rather than the DMPFC usually observed during social cognition (Frith & Frith, 2003). Mitchell has suggested that this occurs because when people try to make sense of similar others, they use the self as a template and project their understanding of themselves onto the similar others.

A self-based explanation cannot account for the MPFC activations typically seen when people make judgments about close others who are not necessarily similar to themselves (van Overwalle, 2009). A recent set of studies by Krienen, Tu, and Buckner (2010) pitted closeness and similarity against one another and found that MPFC activity was more sensitive to closeness than to similarity. When making judgments of a friend whom one acknowledges is not very similar to oneself, robust MPFC activity cannot easily be attributed to a self-projection process.

MPFC has increasingly been associated with empathy processes as well. MPFC activity has been associated with the accuracy of empathic judgments (Zaki, Weber, Bolger, & Ochsner, 2009). Additionally, MPFC activity during an empathy task predicts helping in everyday life (Rameson, Morelli, & Lieberman, 2012). Given that one of the hallmarks of adult empathy is a focus on the needs and experience of the other person rather than on oneself, these findings are hard to reconcile with a "self"-focused account of MPFC.

Finally, MPFC is emerging as a key player in the neural bases of persuasion. Multiple laboratories have now observed MPFC to be more active in response to persuasive messages (Chua et al., 2009; Falk, Berkman, & Lieberman, 2011, in press; Falk, Berkman, Mann, Harrison, & Lieberman, 2010) and in predicting whether people will change their behavior.

Two Theories of MPFC and DMPFC Function

So what does this all mean? What does MPFC really do? In truth, we really don't know yet. Whatever it does will have implications for our understanding of self-processes, including self-knowledge. I have a theory that I think is more than half-baked, but certainly not fully baked. A key fact driving my own theorizing is the asymmetry in MPFC and DMPFC involvement in thinking about the self and thinking about others. As mentioned earlier, when thinking about the self, MPFC is activated in nearly every study, and DMPFC is activated about half the time. In contrast, when thinking about others, DMPFC is activated in nearly every study (91%), and MPFC is reported in one-third (33%) of studies. Thus, it seems that MPFC and DMPFC are not clearly identified as self and social cognition regions per se. It is more plausible to suppose that MPFC is responsible for a mental process that *tends* to be recruited more often when thinking about the self than about others (but it can be involved in either), and that DMPFC is responsible for a mental process that *tends* to be recruited more often when thinking about others than about the self (but it also can be involved in either). So this is our starting point. Any account of the functions of these two regions needs to accommodate that asymmetry.

Generic and Idiosyncratic Theories of People

At least two processing distinctions fit this bill reasonably well. They are related distinctions but are positioned at different levels of analysis. The first and more

straightforward of these is a cognitive distinction between generic and idiosyncratic representations of people. We have a representation of the generic individual, his or her goals and preferences, and how the generic individual is likely to respond in various situations. This is precisely what more than half a century of studies on attribution processes has focused on (Jones et al., 1971). For instance, we would probably all represent the generic person in such a way that we would expect that with a gun pointed to his or her head, the person would feel fear, think about ways to survive or escape this ordeal, and be willing to engage in various low-cost behaviors such as shouting that teen heartthrob "Justin Bieber is my favorite singer of all time," if it would secure freedom. I do not need to know about the person's unique characteristics to make this assessment. It is exactly this sort of generic assessment that theory of mind tasks use repeatedly to measure one person's ability to consider the mental states of another person.

Two important things should be noted about the use of these generic theories of people. First, as social psychologists well know, the fact that people have these theories, use them endlessly, and believe in their utility in no way guarantees that the theories are correct. Second, and more importantly for our current purposes, these same generic theories can be applied to ourselves as well as to others. That is, when I consider what I would do if a gun were held to my head, I might draw on the same generic theory that I use to forecast the reactions of people in general (Karniol, 2003). While we can use this generic theory of people when thinking about others or ourselves, we are probably likely to draw on it more frequently for thinking about others rather than ourselves. For judging others, this generic theory is all we have to go on, but for ourselves we have other kinds of information (Pronin, 2009).

In addition to a generic theory of people, we also have idiosyncratic theories of particular individuals. Idiosyncratic theories no doubt come into play when imagining whether one's father or one's 14-year-old niece would yell "I love Justin Bieber" based solely on a verbal request (without a gun or other threat). Of course, our most idiosyncratic theory of any individual is reserved for ourselves. We have deeply idiosyncratic theories of ourselves. Thus, idiosyncratic theories probably are recruited more often when thinking about the self than about others. At a first approximation, the differential application of generic and idiosyncratic knowledge when thinking about self and others comports well with the ratio of MPFC and DMPFC activations for each target, self and other. This account also accommodates the finding of greater MPFC when thinking about similar others (Mitchell et al., 2005) because this could reflect the projection of one's idiosyncratic self-theory and also accommodate the finding of greater MPFC when thinking about close others (Krienen et al., 2010) who are not similar, as we are likely to have idiosyncratic theories of them as well.

Why would the brain be set up to separate these functions, and what are the implications for our understanding of self-knowledge? One possibility is that generic and idiosyncratic social knowledge functions evolved at different points in our history. A second and more interesting possibility is that there are different computational requirements for each, and that the requirements are sufficiently at odds with one another that it is too computationally costly to try to represent both functions in the same brain region. McClelland, McNaughton, and O'Reilly (1995) gave an elegant demonstration of something analogous in the domain of memory. They were addressing why semantic and episodic memory (i.e., memory for generalities

and memory for specific instances) are represented separately in the brain. They created computational simulations that repeatedly produced "catastrophic interference" when episodic and semantic memories were represented in a single system.

This latter account is exciting because it would suggest that there are likely computational differences involved in idiosyncratic and generic bases of self-knowledge. Given that these two kinds of self-knowledge do not feel phenomenologically different on first pass, this is a case where neuroimaging might help us draw psychological distinctions and suggest avenues of psychological research we might otherwise overlook.

This account would also suggest the possibility that self-knowledge and self-representation more generally might have been an accidental sied effect of needing to represent others in one's group idiosyncratically. Evolutionarily, this might have been the greater press. If this is the case, we should expect that in human children and in other species, the development of idiosyncratic, more so than generic, theories of others would be linked to the development of self-knowledge.

Immersive and Transactional Social Experience

The second possible distinction is phenomenological rather than computational in nature (see Buber [1937] for a similar distinction described from a philosophical perspective and Clark & Mills [1979] on exchange and communal processing which has much in common with the current distinction). Many of our social interactions and the social cognition that supports them are transactional in nature. We are focused on a particular transaction, and other individuals, who happen to have minds that we must take into account, are a means to an end rather than an end in themselves. During these interactions, other people only vaguely rise above the level of other objects or perhaps complex machines that are represented in terms of input–output patterns. At a restaurant, I am aware that if I motion my hand in a certain way the waiter will bring the check over. The waiter is simply a means to an end and this is a two-way street. Many of the canned compliments from people in the service industries do not represent a genuine interest in the customer but rather an understanding of reciprocity and ingratiation that are likely to increase sales and tips. Driving on the road with other cars is probably an ideal example of transactional social cognition. I know that the other cars are being operated by people with minds, and my theory of those minds is integral to how I behave in relation to those other cars, but I am not the least bit interested in those people as ends in themselves.

Transactional social experience is closely aligned with generic theories of people, but here I am focusing on the distinctly diminished sense of human interaction that often parallels the use of this generic theory. Transactional social experience is also not the same as dehumanization (Harris & Fiske, 2006), in that we can and do have transactional experiences with all the people we are closest with in life. Though, to be sure, dehumanization likely increases the tendency to treat others in transactional terms.

While we do not typically treat ourselves transactionally, we certainly can, and our memories of what we did or felt in the past are certainly influenced at times by our generic theories of how the average person would react in that situation. Thus, if DMPFC supports social cognition framed in a transactional way, it would follow that

we would see it often in the kinds of abstract theory of mind tasks typically used in fMRI research and only occasionally when people think about themselves.

Naturally, we do not always treat other people in a transactional manner. Sometimes another person's humanity jumps out at us, and the full appreciation that the person is a sentient being full of hopes, desires, fears, and all the rest captures us. There are times when we really *connect* with another person in a way that has strong emotional and physiological components. This occurs when we empathize or sympathize with what someone else is going through. It probably also happens when actors using the Stanislavski method are fully immersed in the experience of the characters they are playing. And finally, it also happens when we relive, rather than simply recall, our own past experiences. Put another way, we treat ourselves in a far more immersive than transactional way compared with how we treat others, and our degree of connectedness with the other is likely to mediate this variable. Thus, if MPFC supports social cognition that is more immersive in nature, it would follow that we would see it most often when we think about ourselves, a target whose experience we can really dive into, and also in some situations when we think about others.

While the transactional and generic theory of mind accounts line up with one another quite nicely, the immersive and idiosyncratic accounts have an important difference. Certainly it is the case that the more idiosyncratic the knowledge we have of a person, the more easily we can find ourselves in an immersive encounter with that individual. However, it is also the case that we can have an immersive experience with someone about whom we have no idiosyncratic knowledge whatsoever. When we see a starving child on late-night television our understanding is immediate, emotional, and immersive—we are brought into the world of that child—but not based on specialized knowledge we have of the individual. At this point, it is unclear exactly how to reconcile the two accounts, but each focuses on ways of knowing and encountering people (including ourselves) rather than on the distinction between self and other processing per se.

Conclusions?

It is largely out of convention rather than necessity that there is a conclusions section here. The truth is that there are still far more questions than answers about the manner by which the brain supports self-knowledge, and what it has to tell us that is of psychological interest. We know that thinking about oneself, whether about one's autobiographical past, one's trait self-knowledge, or one's current preferences, all activate MPFC extremely reliably, and precuneus$_{PCC}$ and DMPFC somewhat reliably. We know that these are our primary targets for connecting the study of self-knowledge to the brain, that these are the regions where we would expect to see dissociations based on key psychological distinctions within self-knowledge processes and contents. However, what is mostly known is that these are the targets. The further utility of this brain mapping largely awaits further studies from psychologists who find the benefit to using dissociations and convergences at the level of the brain to complement the use of techniques such as self-report, reaction times, and memory clustering.

I have suggested two possible accounts of what MPFC (and DMPFC) may do in a larger, functional sense and how this might relate to how we think about self-

knowledge. Neither of these accounts is fully fleshed out, but they at least suggest a method by which neuroscience may genuinely contribute to the psychological study of self-knowledge. If we can identify the functions of these regions that self-knowledge clearly relies upon, then we may be able to derive additional insights about what is involved in the formation, representation, and retrieval of self-knowledge.

REFERENCES

Beer, J. S., John, O. P., Scabini, D., & Knight, R. T. (2006). Orbitofrontal cortex and social behavior: Integrating self-monitoring and emotion–cognition interactions. *Journal of Cognitive Neuroscience, 18*, 871–879.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 1–62). New York: Academic Press.

Blakemore, S.-J., den Ouden, H., Choudhury, S., & Frith, C. (2007). Adolescent development of the neural circuitry for thinking about intentions. *Social Cognitive and Affective Neuroscience, 2*, 130–139.

Buber, M. (1937). *I and thou*. New York: Scribners & Sons.

Butler, J. (1819). *The analogy of religion*. Hartford, CT: Samuel T. Goodrich. (Original work published 1736)

Clark, M. S., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology, 37*, 12–24.

Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., et al. (1999). In search of the self: A positron emission tomography study. *Psychological Science, 10*, 26–34.

D'Argembeau, A., Stawarczyk, D., Majerus, S., Collette, F., Van der Linden, M., & Salmon, E. (2010). Modulation of medial prefrontal and inferior parietal cortices when thinking about past, present, and future selves. *Social Neuroscience, 5*, 187–200.

D'Argembeau, A., Xue, G., Lu, Z.-L., Van der Linden, M., & Bechara, A. (2008). Neural correlates of envisioning emotional events in the near and far future. *NeuroImage, 40*, 398–407.

Ersner-Hershfield, H., Wimmer, G. E., & Knutson, B. (2009). Saving for the future self: Neural mea sures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience, 4*, 85–92.

Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2011). Neural activity during health messaging predicts reductions in smoking above and beyond self-report. *Health Psychology, 30*, 177–185.

Falk, E. B., Berkman, E. T., & Lieberman, M. D. (in press). From neural responses to population behavior: Neural focus group predicts population level media effects. *Psychological Science*.

Falk, E. B., Berkman, E. T., Mann, T., Harrison, B., & Lieberman, M. D. (2010). Predicting persuasion-induced behavior change from the brain. *Journal of Neuroscience, 30*, 8421–8424.

Farb, N. A. S., Segal, Z. V., Mayberg, H., Bean, J., McKeon, D., Fatima, Z., et al. (2007). Attending to the present: Mindfulness meditation reveals distinct neural modes of self-reference. *Social Cognitive and Affective Neuroscience, 2*, 313–322.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*, 459–473.

Gallup, G. G., Jr. (1970). Chimpanzees: Self-recognition. *Science, 167*, 86–87.

Greenwald, A. G., & Banaji, M. R. (1989). The self as a memory system: Powerful, but ordinary. *Journal of Personality and Social Psychology, 57*, 41–54.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme outgroups. *Psychological Science, 17*, 847–853.

James, W. (1890) *Principles of psychology* (Vol. 1). New York: Holt.

James, W. (1892). *Psychology.* New York: Holt.

Jones, E. E., Kanouse, D. E., Kelley, H. H., Nisbett, R. E., Valins, S., & Weiner, B. (1971). *Attribution: Perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

Karniol, R. (2003). Egocentrism versus protocentrism: The status of self in social prediction. *Psychological Review, 110*, 563–580.

Kelley, W. M. C., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self?: An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*, 785–794.

Klein, S. B., Loftus, J., & Kihlstrom, J. F. (1996). Self-knowledge of an amnesic patient: Toward a neuropsychology of personality and social psychology. *Journal of Experimental Psychology, 125*, 250–260.

Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology, 63*, 739–753.

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*, 13906–13915.

Libby, L. K., & Eibach, R. P. (2002). Looking back in time: self-concept change affects visual perspective in autobiographical memory. *Journal of Personality and Social Psychology, 82*, 167–179.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289.

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). *Handbook of social psychology* (5th ed., pp. 143–193). New York: McGraw-Hill.

Lieberman, M. D., Jarcho, J. M., & Satpute, A. B. (2004). Evidence-based and intuition-based self-knowledge: An fMRI study. *Journal of Personality and Social Psychology, 87*, 421–435.

Locke, J. (1689/1975). *An essay concerning human understanding.* Oxford, UK: Oxford University Press.

Loewenstein, G., & Elster, J. (1992) *Choice over time.* New York: Russell Sage Foundation.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14*, 647–654.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why are there complimentary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419–457.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*, 1306–1315.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*, 655–663.

Moran, J. M., Heatherton, T. F., & Kelley, W. M. (2009). Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience, 4*, 197–211.

Nyberg, L., Kim, A. S. N., Habib, R., Levine, B., & Tulving, E. (2010). Consciousness of subjective time in the brain. *Proceedings of the National Academy of Sciences USA, 107*, 22356–22359.

Pfeifer, J. H., Lieberman, M. D., & Dapretto, M. (2007). "I know you are but what am I?!": An fMRI study of self-knowledge retrieval during childhood. *Journal of Cognitive Neuroscience, 19*, 1323–1337.

Pronin, E. (2009). The introspection illusion. *Advances in Experimental Social Psychology, 41*, 1–67.

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.

Rameson, L. T., Satpute, A. B., & Lieberman, M. D. (2010). The neural correlates of implicit and explicit self-relevant processing. *NeuroImage, 50*, 701–708.

Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.

Schmitz, T. W., Rowley, H. A., Kawahara, T. N., & Johnson, S. C. (2006). Neural correlates of self-evaluative accuracy after traumatic brain injury. *Neuropsychologia, 44*, 762–773.

Schnyer, D. M., Nicholls, L., & Verfaellie, M. (2005). The role of VMPC in metamemorial judgments of content retrievability. *Journal of Cognitive Neuroscience, 17*, 832–846.

Semendeferi, K., Schleicher, A., Zilles, K., Armstrong, E., & Van Hoesen, G. W. (2001). Evolution of the hominoid prefrontal cortex: Imaging and quantitative analysis of area 10. *American Journal of Physical Anthropology, 114*, 224–241.

Semendeferi, K., Teffer, K., Buxhoeveden, D. P., Park, M. S., Bludau, S., Amunts, K., et al. (2011). Spatial organization of neurons in the frontal pole sets humans apart from great apes. *Cerebral Cortex, 21*, 1485–1497.

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

Wang, A. T., Lee, S. S., Sigman, M., & Dapretto, M. (2006). Developmental changes in the neural basis of interpreting communicative intent. *Social Cognitive and Affective Neuroscience, 1*, 107–121.

Way, B. M., Creswell, J. D., Eisenberger, N. I., & Lieberman, M. D. (2010). Dispositional mindfulness and depressive symptomatology: Correlations with limbic and self-referential neural activity during rest. *Emotion, 10*, 12–24.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences, USA, 106*, 11382–11387.

Vazire.indb 76 3/1/2012 11:12:48 AM