Available online at www.sciencedirect.com
SCIENCE DIRECT®
Consciousness and Cognition 14 (2005) 647–660
Consciousness and Cognition
www.elsevier.com/locate/concog

Introduction

Varieties of self-systems worth having

1. Introduction

Just two and a half short millennia after the Greeks issued the injunction "Know thyself," it seems that mankind has finally developed the tools with which to comply. In scientific circles the transition has been swift. Just thirty years ago, Karl Popper and John Eccles published *The Self and Its Brain*, a detailed presentation of their "interactionist," dualistic ontology of mind (Popper & Eccles, 1977). These days, the popular scientific books tell a very different story. Consciousness, free will, and the self are no longer to be treated as topics of open-ended philosophical debate; instead, science will yield the answers. We are nothing but a pack of neurons, whose workings are beginning to be revealed by modern neuroscientific methods.

These are fine, bold sentiments. But not everyone thinks it is that simple. Most agree that cognitive science, in its many manifestations from evolutionary psychology through behavioral experimentation to functional neuroimaging, is telling us a great deal about our inner workings that is interesting and valuable. However, it isn't clear that it is telling us about the nature of consciousness, free will, or the self. Some popularizers of the science of consciousness, who acknowledge this apparent "explanatory gap" (Levine, 2001), suggest we may be able to sneak up on the big prey bit by bit. However, as some philosophers point out, there are reasons to suppose these quarry will forever elude our scientific grasp (Chalmers, 1996). These philosophers might be accused of possessing ulterior motives for their pessimistic view of science. They might be seeking to resist the scientific erosion of their traditional subject matter. It is all the more interesting, then, that the scientific evidence is beginning to suggest a similar conclusion. As we discuss towards the end of this essay, the "cognitive neuroscience of the self" may not tell us about the fundamental nature of the self in a way we hoped or expected.

Indeed, in stark contrast to Popper and Eccles's "The self and its brain," most current scientific work doesn't even aim to explain the self, free will, or consciousness. Some of the most illuminating work has been understood by some as showing that these things don't exist. The scientific topic (hence, the title) of this special issue – and of the interdisciplinary workshop from which it stems, held at Washington University in St. Louis in April 2004—isn't (as it were) the self itself, but rather the cognitive processes which allow us to construct and maintain representations of ourselves. We will return to the issue of the extra-scientific nature of the self near the end of this introduction. The bulk of our discussion will concern itself first with reviewing what we know about the "self-systems" that perform the function of representing various aspects of ourselves, and second with the question of how these systems might have evolved.

We speak of "self-systems worth having" to reflect four characteristics of the recent literature on the self. First, most models imply that the self is supported by a federation of specialized processes rather than a single integrated cognitive function. Second, most researchers think that the phenomenology of selfhood results from the aggregate of the functions performed by these different information-processing devices. Third, most of the information-processing is construed as sub-personal, hence inaccessible to conscious inspection. Fourth, we talk about systems *worth having* to emphasize that there is nothing inevitable about the functioning of any of these systems. Indeed, we survey many cases of pathologies in which these systems are impaired. Further,

1053-8100/$ - see front matter
doi:10.1016/j.concog.2005.08.002

some of these systems stem from species-specific properties of human brains and therefore should be analyzed in terms of the advantages they might afford the organisms that have them.

2. Taxonomies of self-systems

As noted above, one major aspect of the current literature is that self-cognition is construed as a *collection* of distinct processes and representations. This may seem paradoxical in light of the traditional philosophical view of the self as unifier of bundles of experiences. However, the fractionation that is proposed in the neurosciences is orthogonal to the perspectival role of the self in experience. The point of the fragmented models we present here is that self-cognition consists of different functions and requires different types of processing and different knowledge-bases. The main argument for this view relies on neuroimaging data, showing differential brain activation with different tasks, and neuropsychology, showing that different aspects of self-representation can be selectively impaired. To say that different systems are involved in supporting different aspects of self-representation does not mean that these processes are computationally isolated (''encapsulated''). But it does imply that they have a particular input format and that they operate according to specific principles.

Fragmented or fractionated models of self-processes are not really new. They pre-date the current wave of neuroimaging and neuropsychological studies. For instance, Neisser made conceptual and empirical distinctions between five domains of self-knowledge, namely: an *ecological* self, a sense of one's own location in and distinctness from the environment; an *interpersonal self*, a sense of oneself as a locus of emotion and social interaction; an *extended* self, a sense of oneself as an individual existing over time; a *private* self, a sense of oneself as the subject of introspectively accessible experience; and a *conceptual* self, comprising all those representations that constitute a self-image, including representations of one's social role and personal autobiography (Neisser, 1988). These do not map neatly onto the distinct processes identified in the current literature, mostly because current descriptions of the systems engaged are mostly based on neural evidence, either from imaging studies or from clinical studies of selective functional impairment, leading to a different way of parsing the systems involved.

From this cognitive neuroscience standpoint, Gallagher distinguishes broadly between the ''minimal'' and the ''narrative'' self. The former supplies the ecological sense of bodily ownership and agency associated with active behavior, while the latter supports the self-image that associates our identity with various episodes (Gallagher, 2000). The minimal self is what is disrupted when, for instance, patients with asomatagnosia mistake one of their limbs for an external object or patients with schizophrenia mistake their intrusive thoughts for external voices. The narrative self is impaired when amnesia or some other type of psychopathology affects one's connection to past episodes. As we will see below, most authors in the field, including contributors to the present issue, accept this broad distinction between the sense of oneself as acting in and on the environment at a time, on the one hand, and the sense of oneself as a unique individual persisting over time, on the other. But many would also argue for further fractionation, noting behavioral and clinical evidence that each of these distinct senses of self is supported by a variety of cognitive processes.

In the following sections, we explore some of the systems engaged in more detail. Beyond these ecological and narrative self-representations, we will also consider the domain of *social* systems that contribute to self-representation. The self-concept includes notions of social identity or moral status that are not strictly speaking part of one's narrative self (and can be detached from the narrative, as we will see presently). In these social systems, we should also include the capacities for mind-reading and empathy. One may balk at the idea of classifying these social systems as self-processes, since they are engaged in the interpretation of other agents' behavior as well. But this would ignore the important functional overlap between understanding others and understanding oneself. A growing body of empirical evidence suggests that these are engaged in self-cognition, in particular in the interpretation of one's own thought and behavior. Indeed, some neuropsychologists suggest that there may be relatively little that is ''special,'' that is, either anatomically or functionally separable, about self-representations. On the basis of a review of the relevant brain systems, Gillihan and Farah, for instance, propose that some ''physical self'' (that is, ecological or minimal self) systems are probably special in this sense, but that most other self-representation systems are probably not (Gillihan & Farah, 2005). That is, neither their operating principles nor their domain of operation seem clearly distinct from systems engaged in the representation of persons in general.

In the following sections, we review in more detail the state of the art concerning these three domains of self-processes—the agentic self, engaged in action and perception; the memory systems that support autobiographical identity; and the social systems for self-interpretation—before summarizing relevant research on the evolutionary background to self-processes.

3. Agency and ownership

Among the most basic features of persons is that they make things happen. Not only do people make things happen, they think of themselves as the authors of such happenings. This mode of self-representation underlies the "sense of agency," that is, the sense one has of oneself as an actor, a doer. The sense of agency takes a number of different forms (Marcel, 2003). One form consists in representation of one's long-term potential for action, that is, thinking of oneself as capable of a certain range of actions (say, running a mile in five minutes) and incapable of others (running the same distance a minute faster). More commonly, the sense of agency is conceptualized as a transient, in-the-moment type of self-awareness: when performing an action, one represents the action as one's own and under one's control. It seems natural to classify this as an aspect of the ecological or minimal self.

The agentic aspects of the ecological or minimal self should be distinguished from its purely somatic aspects, such as the "sense of ownership" of one's body. Studies of asomatagnosia, a neurological disorder in which the sense of ownership of body parts is selectively impaired, suggest that the capacity for bodily self-recognition is partly supported by circuitry in right parietal cortex (specifically, right supramarginal gyrus) and adjacent white matter (Feinberg, Haber, & Leeds, 1990) as well as right frontal regions (Feinberg & Keenan, 2005, this issue). There is also preliminary evidence from functional neuroimaging that recognizing one's own face draws predominantly on right frontal activity (Feinberg & Keenan, 2005, this issue). Evidence of more precise localization of this capacity, however, is mixed (Gillihan & Farah, 2005).

Like the bodily self, the agentic self may be supported by specific brain networks. In an imaging study comparing schizophrenic patients during and after resolution of delusions of control, in which subjects were asked to move a joystick randomly in four different directions, impairment of agentic self-awareness was correlated with hyper-activation of right parietal cortex (Spence & Driver, 1997). In a complementary study of normal individuals, Farrer et al. compared the brain activity of subjects moving a joystick to control a visually presented cursor (experience of agency condition) with the activity of subjects moving a disconnected joystick and watching the cursor's move under the experimenter's control (violation of agency condition). They found that these two conditions correlated with increased activation in insular and parietal cortices, respectively (Farrer et al., 2004). In a verbal counterpart of this paradigm, in which subjects read text aloud and heard either their own or the experimenter's voice, McGuire et al. found increased activation in lateral temporal cortex, especially on the right, in the latter (incongruent voice) condition (McGuire, Silbersweig, & Frith, 1996).

Cognitive neuroscience has helped to illuminate the sense of agency by advancing our understanding of the neural basis of motor control (Frith, Blakemore, & Wolpert, 2000a; Jeannerod & Decety, 1995). There is substantial evidence that the brain controls action by internally modeling relevant aspects of the agent's body and the environment. In forward (predictive) modeling, for example, motor commands are copied to a system that uses this information to predict the sensory effects of the movement. This enables error correction in advance of actual sensory feedback, which tends to lag well behind actual movement because of neural transduction and processing delays. Since forward modeling is accompanied by attenuation of the sensory effects of actual movements, active (self-initiated) movements are proprioceptively "quiet" relative to passive ones, where forward modeling is absent and no such attenuation occurs. Thus, motor control mechanisms provide cues as to which of one's movements correspond to actions and which do not (Hohwy & Frith, 2004).

Some of the most suggestive work on agentic self-awareness has come from studies of subjects whose sense of agency is impaired as the result of disease or injury. In his contribution to this issue, Chris Frith focuses on schizophrenic subjects with delusions of control. These individuals experience their own actions as alien, that is, as controlled by another agent. Classic reports of such experiences are manifestly bizarre: "My fingers pick up the pen, but I don't control them; what they do is nothing to do with me," and "The force moved my lips. I began to speak. The words were made for me" (Mellor, 1970). Frith had originally explained this phenomenon in terms of patients' failure to monitor their intentions to act: since they are unaware of those intentions, their

subsequent movements come as a surprise to them (Frith, 1987). As Frith notes, this explanation was overturned in light of case studies of anarchic hand (''Doctor Strangelove syndrome''), a neurological impairment in which patients cannot control the purposive movements of one of their hands. On one occasion a patient found himself trying to soap a washcloth with the right hand while the left hand kept putting the soap back in the dish, and on another occasion trying to open a closet with one hand while the other hand closed it (Wegner, 2003). Despite being unaware of their intentions to act, patients with anarchic hand, unlike their counterparts with delusions of control, do not experience their actions as alien. They also try to stop or correct the movements of the wayward hand, whereas patients with delusions of control do not attempt any comparable intervention.

Frith now argues that in both anarchic hand and delusions of control there is a defect in the forward modeling system that impedes the attenuation of proprioceptive feedback, thereby causing patients to experience their movements as passive. The main difference between these disorders, he suggests, lies in the fact that patients with schizophrenia, unlike patients with anarchic hand, tend to perceive agency where there is none. Evidence of this tendency comes from recent studies of patients with schizophrenia, including patients with paranoid delusions, who attribute intentionality to moving shapes in animated sequences that controls describe as intentionality-free (Castelli, Happé, Frith, & Frith, 2000). Frith's hypothesis is especially noteworthy in that it links the sense of agency, understood as a form of self-awareness, with the capacity to think about agency in general.

In the normal case, we experience our voluntary movements as the outcome of our agency. One aspect of that experience is our sense that those movements flow from our conscious intentions to act in particular ways. For instance, one forms the conscious intention to fetch a drink from a nearby vending machine, and that intention sets one's body in motion out the office door and down the hall. This seems like a natural and intuitive way to think of the matter; indeed, it seems to lie at the heart of the commonsense conception of agency. According to Wegner, however, this conception of agency rests on a mistake (Wegner, 2003). The main impetus behind Wegner's skeptical challenge to common sense comes from Libet et al.'s celebrated experiment in which the brain activity of subjects planning and executing a simple action (moving a finger) was monitored with EEG (Libet, 1985). Using this paradigm, Libet and colleagues found evidence that subjects' conscious intention to perform the action, though preceding the action itself, occurred only after an initial burst of brain activity heralding the production of the movement (the ''readiness potential''). The implication was clear: the apparent causal power of conscious intentions to act vis-à-vis the production of action may be merely apparent. Largely on this basis—but also citing cases in which the experience of conscious will breaks down, normal (automatisms, facilitated communication) and otherwise (anarchic hand, delusions of control), as well as experimental manipulations in which subjects project a feeling of causal responsibility onto actions they have not performed (as in the ''helping hands'' paradigm)—Wegner argues that our experience of consciously willing our actions is an illusion. But this hypothesis is controversial, to put it mildly (see Nahmias, 2005, this issue, for a critique; also Jack & Robbins, 2004 and other commentaries in Wegner, 2004).

4. Memory and identity

A great part of the self-model consists in representations of autobiographical facts and episodes, in a description of the past self. This is why there is no coherent description of self-processes without consideration of autobiographical memory (Barclay & Subramaniam, 1987; Klein, Cosmides, Costabile, & Mei, 2002). The latter comprises episodic memories but also a semantic store of facts about one's own past. Experimental and observational studies converge to suggest a model of autobiographical memory where these two components are separate (Conway & Pleydell-Pearce, 2000).

Memory supports self-representations in two different ways. First, it provides representational evidence for our assumption that the self is durable, that every experience is gauged against a background of past episodes and that, by extension, future situations matter. Second, autobiographical memory provides this material in a way that supports another tacit assumption, to the effect that the self is *someone* in particular, with durable or indeed essential qualities that make self different from others. This makes autobiographical memory distinct from episodic memory in general, which does not always support such assumptions. For instance, young children before the age of three develop sophisticated episodic memory (capacity to recall and describe an episode) before they include in their memories the unique perspective of a self having experienced it (Fivush, 1997). The child may not see it as her own experience, or may not construe that aspect of the event as fundamental. Even

later in the preschool years, there seems to be little connection between recall of personal experience and construction of the self-concept (Howe & Courage, 1997; Nelson & Fivush, 2004).

Autobiographical memory also contributes to self-coherence by presenting recalled episodes and current properties of the self as mutually explanatory. Typical autobiographical musings or discourse go back and forth between facts and episodes (Thorne, 2000). These combinations generally obey strict narrative constraints, so that past experiences have effects on subsequent behaviors, intentions, and beliefs (Brown & Schopflocher, 1998). Moreover, the narratives are generally framed by a cultural model of the stages normatively identified as typical development, and of the ways in which going through these different stages creates a distinct self (Berntsen & Rubin, 2004).

These findings, together with the vast literature from social and cognitive psychology documenting false memory, memory distortions, self-advantageous biases and resistance to cognitive dissonance (Schacter, Coyle, & Harvard Center for the Study of Mind Brain and Behavior, 1995), would suggest that relevant material is selected (and distorted) from a large store of available episodes. More radically, some psychologists suggest that there is no stable representation of the autobiographical past, so that personal memories are constructed on the hoof on the basis of current self-relevant goals that direct activation of episodes and autobiographical knowledge (Conway & Pleydell-Pearce, 2000).

This view of autobiographical memory as coordinated activation of two systems—rather than a single proprietary system—may explain why the evidence for specific neural correlates is ambiguous, despite important studies (Conway et al., 1999; Fink et al., 1996; Maguire & Mummery, 1999). To date these studies have yielded inconsistent evidence, possibly because they failed to control for potential confounds between autobiographical and other aspects of memory (Gillihan & Farah, 2005). Studies of self-trait memory (Johnson et al., 2002; Kelley et al., 2002; Kjaer, Nowak, & Lou, 2002) suffer from analogous limitations, though several of these studies suggest a special role for medial prefrontal cortex (Gusnard, 2005, this issue).

The self-concept includes a summary description of one's own personality that is crucially dependent on memory. Indirect evidence for this (literally) ''self-serving'' function of memory is that individuals with a more specific description of their own personality also have more specific and easily accessible memories (Neimeyer & Rareshide, 1991). Self-assessments of personality use dimensions and concepts and illustrative episodes that are not formally different from those used in the description of other persons. There is now good evidence that the *summary* description of personality is distinct from the store of *episodes* that illustrate it (Klein, Babey, & Sherman, 1997; Klein, Chan, & Loftus, 1999). Indeed, for knowledge of other people it would seem that activating trait-summary knowledge primes trait-*inconsistent* episodes (i.e., exceptions to what might be expected given the trait), perhaps as a useful limit on inferences from the trait summary (Klein, Cosmides, Tooby, & Chance, 2001). Also, the neuropsychological literature includes dramatic cases of dissociation, in which patients who have lost all episodic memories still preserve an accurate self-description in terms of personality summary (Klein, Rozendal, & Cosmides, 2002).

If memory systems are actively engaged in construction of the self-concept, we should find that many pathologies that affect episodic memory more than semantic stores would correlate with a damaged sense and experience of self. This is indeed and obviously the case in most amnesias, but also in a variety of pathologies such as frontal-lobe damage, autism, and schizophrenia. In all these conditions, the impaired access to episodic stores results in a diminished or fuzzy self-image, with great difficult in representing particular features of one's experience (Klein, 2001). Conversely, we should observe that pathologies that affect the emotional valence of experience correlate with poor access to autobiographical memory. This is the case in depression and post-traumatic stress disorder. Clinically depressed individuals generally have overly generic autobiographical memories, with poor retrieval of what made particular experiences unique and a tendency to focus on repeated scripts (Kuyken & Dalgleish, 1995; Moffitt, Singer, Nelligan, & Carlson, 1994). Trauma victims also report specific problems of autobiographical memory, combining intrusive recollection of the traumatic event with difficulty in voluntary recall (Golier, Yehuda, & Southwick, 1997).

5. Mind-reading and empathy

As we mentioned above, capacities for mind-reading and empathy are involved both in representing other agents' behaviors and making sense of one's own, as well as supporting a distinct self.

Specific capacities for mind-reading or "theory of mind" are geared to interpreting other agents' (or one's own) behavior in terms of goals, beliefs, memories, and inferences (Leslie, 1987; Perner, 1991; Whiten, 1991). For a long time, developmental and cognitive psychologists have debated the extent to which mind-reading is based on *simulation* of other agents' behaviors, on some form of imaginary projection of oneself into their situation (Gordon & Olson, 1998; Nichols, Stich, Leslie, & Klein, 1996). The alternative is to consider mind-reading as the outcome of general-purpose theory-building (Gopnik & Wellmann, 1994) or as a specialized inferential capacity (Leslie, 1994). Even in these models, however, some form of imaginary or off-line engagement of one's own systems is a common by-product of making sense of others' behavior (Nichols & Stich, 2003).

This is most salient in the processes engaged in empathy, in the detection of and reactions to other's feelings and sensations. Empathy develops early, as even 2- and 3-year olds engage in appropriate comforting behaviors when others manifest pain or sadness (Gibbs, 2003). Neuro-imaging studies demonstrate the overlap between circuits engaged, e.g., in the affective states resulting from own pain and representation of other people's pain (Singer et al., 2004) or representation of own and other people's feelings, although other networks, particularly in the right parietal cortex, are specifically engaged in the "own situation" representations (see review in Decety & Sommerville, 2003).

Is empathy a subset or rudimentary form of simulation for mind-reading? Decety and colleagues point out that empathy requires three distinct processes: the vicarious experience of another person's feelings, a "tagging" process that distinguishes this from own experience, and finally some understanding of the causal processes that lead to the experienced feelings (Decety & Jackson, 2004). This functional distinction between empathy and more "cognitive" forms of simulation is confirmed by their dissociation in pathology. As Blair and others have demonstrated, psychopaths have standard mind-reading abilities. By contrast, their capacities for empathy are impaired. The neuro-imaging literature supports this interpretation and differentiates in this sense psychopaths from autistic patients (Blair, 2003). In this issue, Blair expands on this interpretation of psychopathy. Psychopaths differ from other criminals in their lack of remorse and their difficulty in achieving normal control of behavior. Specific impairment of amygdala and orbitofrontal function may result in the specific pathology.

Mind-reading and empathy contribute to self-understanding in several different ways. First, our understanding and explicit representation of our own behaviors require interpretations that are as "theoretical" and inferentially complex as those directed at others. Indeed, the notion that simulation of others' thoughts works because of a privileged self-knowledge and direct access to the causes of one's mental states may be an illusion (Gopnik, 1993). As a limiting-case, consider those intrusive thoughts and unplanned actions that are fairly common in normal experience. The sense of ownership (this thought, this action are *mine*) despite the absence of deliberate will requires complex interpretation (or post hoc rationalisation) linking the thoughts and actions to a background of beliefs and dispositions. Indeed, impairment of the capacity to produce such self-interpretations may well contribute to delusions that other agents are talking in one's head or driving one's movements (Frith, 1999; Frith, Blakemore, & Wolpert, 2000b). Second, it seems that people whose mind-reading capacities are impaired have difficulty representing a distinct self (but see Nichols & Stich, 2003 for a dissenting view). Autistic patients have impaired access to episodic memories that connect to self-representations (Klein et al., 1999; Klein, Cosmides, et al., 2002). Third, the mind-reading literature shows that various empathic and inferential capacities are required to sway our interpretations of current situations away from their natural egocentric bias. A recurrent result in social psychological experiments is the privilege of one's own perspective in the representation of the social world. For instance, we know that most people tend to overestimate the extent to which their actions are noticed by others (Gilovich, Medvec, & Savitsky, 2000) and that they attribute to others their own perspective more than they do the opposite. Representations of the social world tend to be egocentric by default, and to adopt other perspectives is a more effortful and generally deliberate operation. To some extent, mild forms of autism, such as Asperger's syndrome, may be seen as a pathological form of this bias, as Uta Frith and Frederique de Vignemont argue in their contribution to this issue (see also Frith, 2004).

In recent years, considerable attention has been paid to the neural correlates of mind-reading, including the introspective (first-person) variety. Vogeley et al. compared brain activity of subjects reading and answering questions about scenarios involving other persons' mental states, on the one hand, versus scenarios involving

one's own mental states, on the other. They found that while both conditions activated right prefrontal areas, the self condition differentially activated right temporo-parietal junction, anterior cingulate cortex, and bilateral precuneus (Vogeley et al., 2001). Evidence from imaging studies of emotional self-monitoring has been mixed, pointing either to increased activation of anterior cingulate cortex (Lane, Fink, Chau, & Dolan, 1997) or medial prefrontal cortex plus frontal operculum/left insula (Gusnard, Akbudak, Shulman, & Raichle, 2001). Most of these studies suffer from methodological limitations so the results need to be taken with a grain of salt (Gillihan & Farah, 2005). In particular, it would be premature to conclude that the capacity to read one's own mind is anatomically distinct from the capacity to read minds in general, tempting as that thought might seem.

Actual social interaction does not just require the fundamental concepts of theory of mind, but also a vast knowledge-base about the way persons behave, an intuitive human psychology in the broader sense (Astington, 2003). Smooth social interaction generally requires adequate understandings of motivation, feeling, memory, emotion, and reasoning in other agents. Even the simplest operations of everyday mind-reading require assumptions about mental functioning, the connections between intentions and actions, the way people estimate their own actions, the way they are motivated by greed or benevolence or spite, the way a disappointing experience can modify their behavior, and so on. Each of these assumptions is also modulated, in the case of persons we know, by specific parameters that constitute that person's personality.

As Malle notes, this domain of psychology has not really been explored in studies of theory of mind (Malle, 2004). These are generally pitched at the level of foundational concepts (belief, intention) rather than their actual deployment in people's explanations of behavior (Malle, Moses, & Baldwin, 2001). People routinely construct social-psychological and personality-based explanations of others, folk understandings of behavior that are not the object of any coherent theoretical framework. The intuitive explanation of behavior used to be described in social psychology in terms of attribution theory (Heider, 1958), which is psychologically insufficient (Harvey, Town, & Yarkin, 1981) and inconsistent with more recent findings on theory of mind (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000).

This enriched theory of mind or "personology" (Gilbert, 1998) is crucially involved in our perception of others but also in our estimates of the way they perceive us. In this quasi-theoretical understanding of self as a set of dispositions, the egocentric bias also predicts a systematic mismatch between self-image and others' perception. This is explored in this issue by Tom Oltmanns and colleagues in the context of personality disorders. To the extent that there is a common conceptual system describing own and reflected views of personality, there is also a systematic mismatch between these perspectives (Oltmanns, Gleason, Klonsky, & Turkheimer, 2005, this issue).

6. Evolutionary background

Just as for other specialized cognitive systems, it makes sense to ask how the information-processing functions involved in self-representation might have evolved by natural selection (Cosmides & Tooby, 1994); and how their neural implementation would be consistent with such evolutionary explanations (Duchaine, Cosmides, & Tooby, 2001). The adaptationist viewpoint would evaluate the extent to which various aspects of self-systems could be interpreted as solutions to recurrent problems in ancestral environments, or as exaptations (adaptive by-products of previous cognitive adaptations) or as spandrels, non-functional features (see Buss, Haselton, Shackelford, Bleske, & Wakefield, 1998 for these distinctions). Given the complexity and variety of systems involved in self-representation, different aspects of the self will likely fall into each of these categories—although, given the clearly functional nature of most aspects of self-representation, we might expect the latter category, the spandrels, to be the least well represented. In the past, applying the logic of adaptationism to such central and seemingly unique human capacities has often triggered strong resistance. Wallace himself, although the co-creator of natural selection theory, considered self-consciousness as too complex to be one of its outcomes (Wallace, 1889). Note that his main argument was that the sense of self seemed to constitute a radical departure from other forms of phenomenal awareness. But this argument itself relied on the assumption that there is an *integral* self-system. Given that assumption, it seems indeed difficult to consider the self as the result of a slow, incremental process of natural selection, each step of which is conducive to better reproductive potential. It is by contrast more tractable to evaluate the potential evolutionary background of separate self-relevant systems.

From the phylogenetic viewpoint, it seems that some aspects of the ecological or minimal self may have been a precursor to more sophisticated self-representations. The most developed account comes from Povinelli and colleagues (Barth, Povinelli, & Cant, 2004). They consider the puzzling pattern of results in the famous Gallup mirror experiments (Gallup, 1979, 1992, 1994). Experimenters leave a colored dot on the forehead of a sleeping subject who has been trained in the use of mirrors. The expectation is that ownership of one's body image will be manifest in attempts to explore or remove this dot once seen in the mirror. All human children pass the test around the age of 3, some (young) chimpanzees pass the test but not their elders (de Veer, Gallup, Theall, van den Bos, & Povinelli, 2003), and most gorillas fail while most orangutans pass (Gallup et al., 1995). According to Povinelli and colleagues, one possibility is that a sense of the embodiment of self—as opposed to mere proprioception—a sense of ownership of one's own body, may have evolved in some primates as a consequence of arboreal locomotion (Barth et al., 2004). Orangutans need subtle appreciation of their own body position, posture, and weight to brachiate and support themselves on flimsy branches. It is not as though they can navigate by trial and error, since a fall will likely prove fatal. The behavior and the required capacity are less developed in chimpanzees and even less in gorillas. This would suggest a complicated history for this kind of self-representation, having been lost by the primate branch that led to chimpanzees, and developed in the hominine lineage.

Representational systems allow the organism to create a flexible internal model that incorporates the critical variables involved in a task. The advantage is that this allows for accurate generalization to the numerous novel configurations of variables that are encountered in complex problem spaces—situations where trial and error learning would prove unreliable if possible at all. The costs are those of maintaining greater computational resources, and longer processing time to compute a suitable action. Whilst no detailed accounts exist for aspects of the self other than body image, we can attempt to outline a plausible trajectory, bearing in mind the costs and advantages of these systems. Perhaps the simplest next step is to seek to explain the other aspect of the minimal self, the ability to represent one's own actions and their outcomes. This provides a database that is essential to planning sequences of actions in order to achieve a specific outcome. However, these abilities would need to have evolved as add-ons to more simple mechanisms for selecting and initiating actions, and there would be reasons for maintaining these simpler systems, which are better suited to rapid and repetitive tasks. Hence our system for representing our own actions is at least partially independent of the mechanisms that select and initiate many of our actions—hardly surprising from an evolutionary point of view, although the resulting errors can seem counterintuitive from the first person phenomenological perspective (Wegner, 2003).

A further issue concerns why our representations of our own action should be explicit or "declarative," in the sense that they are available for verbal description, instead of remaining encapsulated within a module dedicated to deliberative action. Humans are adapted to the specific ecology of what physical anthropologists call the "cognitive niche" (Tooby & DeVore, 1987). In this setting, a critical capacity consists in the ability to provide and understand task instructions. This would help impart survival skills to kin, facilitate cooperation in group tasks such as hunting, and more generally it allows the individual to maximize their gain from a single learning episode: once by acquiring new knowledge for themselves, and on further occasions by bartering that knowledge in exchange for other knowledge or resources. The corresponding increase in the quantity and speed with which individuals can acquire knowledge, relative to what can be acquired through individual learning, may enable, for instance, rapid adaptation to unfamiliar environments.

Conspecifics do not simply cooperate to gain resources, they also compete for resources. The question of how competitive and cooperate dynamics may have combined is essential to understanding some aspects of the self. An additional level of complexity is added by the question of how the capacity to model the self relates to the capacity to model others. As Menander, a third century B. C. Greek dramatist, put it, "'Know thyself' is a good saying, but not in all situations. In many it is better to say 'know others.'" In the social context, the individual is faced with important decisions concerning who they might most profitably cooperate with, as well as who they can profitably deceive (Cosmides & Tooby, 1992). Given these conditions, cognitive capacities involved in social interaction are likely to become more complex and fine-grained, as even small progress in the monitoring of socially transmitted information could be highly beneficial (Dunbar, 2003; Povinelli & Prince, 1998). If, as we suggested earlier, systems for understanding ourselves and others are overlapping, this may be because some self-systems evolved in tandem with our capacity to understand others. Given the "un-

seen'' variables involved in predicting behavior, it may be that internal access to information about a similar system proved very useful. This view is closely allied to ''simulation theory.'' If it is correct, then selective pressure to understand others would have in turn put pressure on better self-understanding. However, we expect that the picture will ultimately prove far more complex, as there is still much to understand about the functional and neural signatures of these processes.

Whether or not selection for the ability to understand others may have directly resulted in selective pressure for self-understanding, it will have placed indirect pressure by changing the environment. Once the fitness of individuals depends upon whether they are selected for cooperation by others, it becomes valuable to be able to convince others that one is a trustworthy cooperator. The most obvious immediate adaptation to suit this purpose would be our tendency to rationalize—our tendency to provide (largely positive) explanations for our actions, even though we often lack knowledge of their true causes (Wegner, 2003). This dynamic may also help to explain a rather more mysterious feature of the self: the narrative self. Why do we seem to care so much that we can construct a (positive) narrative of our lives?

7. Evolving the narrative self

Consider the sense of autobiographical coherence of the self. As Conway argues on the basis of pathology and normal function, autobiographical memories are mostly accessed and modified in order to maintain a coherent self narrative (Conway & Pleydell-Pearce, 2000). This means not just a coherent story made of re-experienced episodes, but a set of episodes that can match current self-model and goals (Conway, 1992). But why should historical coherence matter? Studies of autobiographical memory reveal that memory is to some extent ''self-serving''—this at least is the more attention-grabbing part of the experimental results. However, the studies also reveal that the possible range of memory revision is actually rather limited, not just by plausibility but also by the resistance of memory material, and by the gulf in saliency between experienced and generated material (see studies in Schacter et al., 1995). In other words, memory is self-serving but only to the extent that self-relevant goals remain, by and large, memory-compatible. What we seem to be doing is producing a good sales description for our character—never knowingly dishonest, but nonetheless the most positive possible evaluation of the facts. By preparing a ready narrative, and sharing it with others, we can go a step further than merely rationalizing our actions ad hoc. The problem with ad hoc rationalizations is, of course, that over a period of time others will discover their inconsistency—at which point they will have a deleterious rather than a positive effect on our perceived trustworthiness. The narrative self helps to make the rationalizations we offer to others consistent.

This account seems plausible when we consider the emphasis we place on character traits in our evaluations of others. We assume these traits are critical determinants of morally relevant aspects of behavior. However, social and cognitive psychologists have produced a mountain of empirical evidence that situation rather than character is the main predictor of behavior (Doris, 2002). Why then do we cling to this belief in the central role of character? A related question is: If the narrative self is never anything more than a fabricated sales description, devoid of fact, why should anyone pay any attention to it? If character traits lack all predictive value, we should surely have evolved not to consider them as important. Further, since self narratives can be quite effortful to create and maintain, we should have evolved to no longer invest resources in them. Why, then, do people seem to care so much about their personal narrative, as well as the narratives of others? One plausible solution to this puzzle is that our personal narratives are predictive of some important aspects of behavior, we don't possess stable moral character traits. Experimental work shows that character traits do little to predict behavior in certain situations, such as when an authority figure instructs us to inflict harm on another. However, the false belief that acts of this sort are indicative of character can be self-sustaining, influencing the individual's behavior in other settings, where behavior is less constrained by the situation. An extreme example is a strategy used to recruit child soldiers in West Africa. The child is kidnapped, given a gun and instructed, under threat of being shot themselves, to shoot a member of their own village. Having committed this act, the children are convinced that they must stay with the army and continue in further violence, since their village would never take them back.

The narrative self illustrates two important features of the way we represent ourselves. The first is that it is in some way irrational. In this case, our view of character is, at least in part, simply false. Almost everyone

believes they wouldn't give electric shocks to an unknown innocent just because an authority figure tells them to. Yet the empirical data are quite clear that one will. The second important feature is the importance of self-identification. In this case, once we identify ourselves as having particular character traits, we will work hard to maintain consistency between our actions and these traits. Nietzsche seemed to understand these points when he remarked: "Active, successful natures act, not according to the dictum 'know thyself,' but as if there hovered before them the commandment: will a self and thou shalt become a self" (Nietzsche, 1879/1986).

8. Irrational Identification

Another seemingly irrational facet of our behavior is our tendency to respond to others' distress with empathy, what Suzanne Langer described as an "involuntary breach of selfhood" (Langer, 1948). We seek to reduce the suffering of others as if it were our own, even when doing so depletes our resources to no apparent advantage. This tendency may have evolved to facilitate effective parenting, yet it is less easy to see why it should be applied to unrelated individuals. Again, the function may be to signal trustworthiness. This line of reasoning is parallel to a set of evolutionary psychology arguments based on strategic modeling (Tooby & DeVore, 1987). What matters here is that particular beliefs could compel us to adopt strategies that are not immediately beneficial, but demonstrate a will to cooperate rather than defect (Schelling, 1960). Economists have shown how such strategies can evolve and lead to non-opportunistic equilibrium (McCabe & Smith, 2001). In the moral domain, Robert Frank argued that moral feelings may play a similar role. To the extent that these feelings create a motivation for cooperative behavior *and* that they are manifestly beyond voluntary control, they constitute honest signals that are in the long term beneficial to individual fitness (Frank, 1988) (see also Gintis, 2000 for a more recent appraisal.

Empathy, however, remains particularly difficult to explain, not least because we do seem to be able to pick and choose who take as targets of concern. There even appear to be well-established social mechanisms for this, such that expressions of disgust and contempt are used to label individuals as something less than persons, "mere bodies without souls," toward whom we need feel no moral duty or concern (Bloom, 2004). Why, then, don't we always pick and choose when it would be to our long-term advantage to show empathy versus acting ruthlessly? Again, it seems there must be a powerful effect of self-identification. Something about our make up, presumably, makes it very hard for us both to maintain the appearance of allegiance to a group and to act ruthlessly towards members of the group when we find this to our advantage.

9. A new folk-psychology of self?

Recent developments in the cognitive studies may have an effect, not just on our scientific explanations of self-relevant cognitive systems, but also on the general, folk-scientific notion of what a self is. There are many complex interactions between scientific developments and our folk-psychology, especially in its explicit, normative and historically specific manifestations. Although the sciences of the mind strive to identify genuine natural kinds of processes, functions and dysfunctions, they also generate what Hacking called "interactive kinds," that is, they create frames for people's folk-understandings, which often in turn shape scientific questions about cognition and psychopathology (Hacking, 1995a). The development of nosological categories like "multiple personality" is best seen as a set of reciprocal influences between social movements, popular understandings, and scientific agendas (Hacking, 1995b, 1998).

In this vein, it may be of interest to ask to what extent neurocognitive understandings of the self might influence folk-theories. For most cognitive scientists, the popularization of such themes as the modularity of brain systems, their dependence on chemical communication, the effects of their impairment on personal experience, is and could not but be a Good Thing. From this perspective, the more widespread science is, the more effectively it can dispel superstition or vague folk-understandings. And so it does—up to a point. A problem in this case, however, is that the claims of cognitive science may be incommensurable with the questions that prompt folk-understandings. In this sense, the claim that "we are nothing but a pack of neurons" may be seen as irrelevant or inflammatory–not because it is metaphysically wrong, but because the statement works as an injunction to revise our self model without providing any useful guidance on how to do this. Instead, by merely emphasizing our physical nature without putting the equation to any useful effect, all the statement succeeds

in doing is tweaking a tendency to treat people with less moral concern (Bloom, 2004). This creates two problems. First, it is likely to be socially divisive (Farah, 2005a, 2005b). Second, it is liable to mislead, because it involves a spurious claim to authority. Science may deliver the definitive story about natural kinds, but it cannot claim the same authority about interactive kinds. In other words, science may tell us exactly what we are made of, but it cannot tell us who we really are. We construct our own selves, based in good part on whom and what we identify with. Science can influence and guide this process, by telling us about the workings of the mechanisms that support self-representation. In doing so it may help us to improve our self-models in a variety of ways—making those models more efficient, reducing the scope and severity of mental illness, helping to promote social cohesion—but science cannot dictate how we should think of ourselves. Further work by both scientists and philosophers will be needed to understand how the science of self-systems can effectively feed into renewed folk-understandings of the self.

In the meantime, the studies presented in this special issue demonstrate that significant and rapid progress has been made in understanding the processes underlying self-representation. In all likelihood this progress will continue apace. As human beings we can all agree, with Popper and Eccles, that our selves have a brain; as scientists and philosophers we are just beginning to understand why our brains need selves.

References

Astington, J. W. (2003). Sometimes necessary, never sufficient: False-belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 13–38). New York: Psychology Press.

Barclay, C. R., & Subramaniam, G. (1987). Autobiographical memories and self-schemata. *Applied Cognitive Psychology, 1*(3), 169–182.

Barth, J., Povinelli, D. J., & Cant, J. G. H. (2004). Bodily origins of self. In D. R. Beike, J. M. Lampinen, et al. (Eds.), *The self and memory* (pp. 11–43). New York: Psychology Press.

Berntsen, D., & Rubin, D. C. (2004). Cultural life scripts structure recall from autobiographical memory. *Memory & Cognition, 32*(3), 427–442.

Blair, R. J. R. (2003). Neurobiological basis of psychopathy. *British Journal of Psychiatry, 182*(1), 5–7.

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

Brown, N. R., & Schopflocher, D. (1998). Event clusters: An organization of personal events in autobiographical memory. *Psychological Science, 9*(6), 470–475.

Buss, D. M., Haselton, M. G., Shackelford, T. K., Bleske, A. L., & Wakefield, J. C. (1998). Adaptations, exaptations, and spandrels. *The American Psychologist, 53*(5), 533–548.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage, 12*(3), 314–325.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Conway, M. A. (1992). *Theoretical perspectives on autobiographical memory*. Dordrecht: Kluwer Academic Publishers.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107*(2), 261–288.

Conway, M. A., Turk, D. J., Miller, S. L., Logan, J., Nebes, R. D., Meltzer, C. C., et al. (1999). A positron emission tomography (PET) study of autobiographical memory retrieval. *Memory, 7*(5–6), 679–702.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 85–116). New York: Cambridge University Press.

de Veer, M. W., Gallup, G. G., Jr., Theall, L. A., van den Bos, R., & Povinelli, D. J. (2003). An 8-year longitudinal study of mirror self-recognition in chimpanzees (pan troglodytes). *Neuropsychologia, 41*(2), 229–234.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 406–412.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences, 7*(12), 527–533.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. New York: Cambridge University Press.

Duchaine, B., Cosmides, L., & Tooby, J. (2001). Evolutionary psychology and the brain. *Current Opinion in Neurobiology, 11*(2), 225–230.

Dunbar, R. (2003). Evolution of the social brain. *Science, 302*(5648), 1160–1161.

Farah, M. J. (2005a). Neuroethics: The practical and the philosophical. *Trends in Cognitive Sciences, 9*(1), 34–40.

Farah, M. J. (2005b). Reply to Jedlicka: Neuroethics, reductionism and dualism. *Trends in Cognitive Sciences, 9*(4), 173.

Farrer, C., Franck, N., Frith, C. D., Decety, J., Georgieff, N., d'Amato, T., et al. (2004). Neural correlates of action attribution in schizophrenia. *Psychiatry Research: Neuroimaging, 131*(1), 31–44.

Feinberg, T. E., Haber, L. D., & Leeds, N. E. (1990). Verbal asomatognosia. *Neurology, 40*(9), 1391–1394.

Feinberg, T., & Keenan, J. P. (2005). Where in the brain is the self? *Consciousness and Cognition, 14*(4).
Fink, G. R., Halligan, P. W., Marshall, J. C., Frith, C. D., Frackowiak, R. S. J., & Dolan, R. J. (1996). Where in the brain does visual attention select the forest and the trees? *Nature, 382*(6592), 626–628.
Fivush, R. (1997). Event memory in early childhood. In N. Cowan (Ed.), *The development of memory in childhood* (pp. 139–161). Hove, UK: Psychology Press.
Frank, R. (1988). *Passions within reason. The strategic role of the emotions*. New York: Norton.
Frith, C. D. (1987). The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. *Psychological Medicine, 17*(3), 631–648.
Frith, C. D. (1999). How hallucinations make themselves heard. *Neuron, 22*(3), 414–415.
Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000a). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B Biological Sciences, 355*(1404), 1771–1788.
Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000b). Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action. *Brain Research Reviews, 31*(2–3), 357–363.
Frith, U. (2004). Emanuel Miller lecture: Confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry, 45*(4), 672–686.
Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14–21.
Gallup, G. G. (1979). Self-awareness in primates. *American Scientist, 67*(4), 417–421.
Gallup, G. G. (1992). Levels, limits, and precursors to self-recognition: Does ontogeny recapitulate phylogeny? *Psychological Inquiry, 3*(2), 117–118.
Gallup, G. G. (1994). Monkeys, mirrors, and minds. *Behavioral and Brain Sciences, 17*(3), 572–573.
Gallup, G. G., Povinelli, D. J., Suarez, S. D., Anderson, J. R., Lethmate, J., & Menzel, E. W. Jr., (1995). Further reflections on self-recognition in primates. *Animal Behaviour, 50*(6), 1525–1532.
Gibbs, J. C. (2003). *Moral development and reality: Beyond the theories of Kohlberg and Hoffman*. London: Sage Publications.
Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, et al. (Eds.), *The handbook of social psychology* (Vol. 2, 4th ed., pp. 890–150). New York: Oxford University Press.
Gillihan, S. J., & Farah, M. J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin, 131*(1), 76–97.
Gilovich, T., Medvec, V. H., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology, 78*(2), 211–222.
Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169–179.
Golier, J. A., Yehuda, R., & Southwick, S. M. (1997). Memory and posttraumatic stress disorder. In P. S. Appelbaum & L. A. Uyehara, et al. (Eds.), *Trauma and memory: Clinical and legal controversies* (pp. 225–242). New York: Oxford University Press.
Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Brain and Behavioral Sciences, 16*, 1–14.
Gopnik, A., & Wellmann, H. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain-specificity in cognition and culture*. New York: Cambridge University Press.
Gordon, A. C. L., & Olson, D. R. (1998). The relation between acquisition of a theory of mind and the capacity to hold in mind. *Journal of Experimental Child Psychology, 68*(1), 70–83.
Gusnard, D. (2005). Being a self: Considerations from functional imaging. *Consciousness and Cognition, 14*(4).
Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(7), 4259–4264.
Hacking, I. (1995a). The looping effects of human kinds. In D. Sperber, D. Premack, et al. (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). New York: Oxford University Press.
Hacking, I. (1995b). *Rewriting the soul: Multiple personality and the sciences of memory*. Princeton, NJ: Princeton University Press.
Hacking, I. (1998). *Mad travelers: Reflections on the reality of transient mental illnesses*. Charlottesville, VA: University Press of Virginia.
Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is "the fundamental attribution error"? *Journal of Personality and Social Psychology, 40*(2), 346–349.
Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
Hohwy, J., & Frith, C. (2004). Can neuroscience explain consciousness? *Journal of Consciousness Studies, 11*(7–8), 180–198.
Howe, M. L., & Courage, M. L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104*(3), 499–523.
Jack, A. I., & Robbins, P. (2004). The illusory triumph of machine over mind: Wegner's eliminativism and the real promise of psychology. *Behavioral and Brain Sciences, 27*(5), 665–666.
Jeannerod, M., & Decety, J. (1995). Mental motor imagery: A window into the representational stages of action. *Current Opinion in Neurobiology, 5*(6), 727–732.
Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain, 125*, 1808–1814.
Kelley, W. M., 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*(5), 785–794.
Kjaer, T. W., Nowak, M., & Lou, H. C. (2002). Reflective self-awareness and conscious states: Pet evidence for a common midline parietofrontal core. *Neuroimage, 17*(2), 1080–1086.

Klein, S. B. (2001). A self to remember: A cognitive neuropsychological perspective on how self creates memory and memory creates self. In C. Sedikides & M. B. Brewer (Eds.), *Individual self, relational self, collective self* (pp. 25–46). Philadelphia: Psychology Press.

Klein, S. B., Babey, S. H., & Sherman, J. W. (1997). The functional independence of trait and behavioral self-knowledge: Methodological considerations and new empirical findings. *Social Cognition, 15*(3), 183–203.

Klein, S. B., Chan, R. L., & Loftus, J. (1999). Independence of episodic and semantic self-knowledge: The case from autism. *Social Cognition, 17*(4), 413–436.

Klein, S. B., Cosmides, L., Costabile, K. A., & Mei, L. (2002). Is there something special about the self? A neuropsychological case study. *Journal of Research in Personality, 36*(5), 490–506.

Klein, S. B., Rozendal, K., & Cosmides, L. (2002). A social-cognitive neuroscience analysis of the self. *Social Cognition, 20*(2), 105–135.

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2001). Priming exceptions: A test of the scope hypothesis in naturalistic trait judgments. *Social Cognition, 19*(4), 443–468.

Kuyken, W., & Dalgleish, T. (1995). Autobiographical memory and depression. *British Journal of Clinical Psychology, 34*(1), 89–92.

Lane, R. D., Fink, G. R., Chau, P. M., & Dolan, R. J. (1997). Neural activation during selective attention to subjective emotional responses. *Neuroreport, 8*(18), 3969–3972.

Langer, S. K. K. (1948). *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. New York: Penguin.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind". *Psychological Review, 94*, 412–426.

Leslie, A. M. (1994). ToMM, ToBY and agency. Core architecture and domain-specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain-specificity in cognition and culture* (pp. 119–148). New York: Cambridge University Press.

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. New York: Oxford University Press.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), 529–566.

Maguire, E. A., & Mummery, C. J. (1999). Differential modulation of a common memory retrieval network revealed by positron emission tomography. *Hippocampus, 9*(1), 54–61.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology, 79*(3), 309–326.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (Eds.). (2001). *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.

Marcel, A. J. (2003). Introspective report: Trust, self-knowledge and science. *Journal of Consciousness Studies, 10*(9–10), 167–186.

McCabe, K. A., & Smith, V. L. (2001). Goodwill accounting and the process of exchange. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 319–340). Cambridge, MA: MIT Press.

McGuire, P. K., Silbersweig, D. A., & Frith, C. D. (1996). Functional neuroanatomy of verbal self-monitoring. *Brain, 119*, 907–917.

Mellor, C. S. (1970). First rank symptoms of schizophrenia. *British Journal of Psychiatry, 117*(536), 15–23.

Moffitt, K. H., Singer, J. A., Nelligan, D. W., & Carlson, M. A. (1994). Depression and memory narrative type. *Journal of Abnormal Psychology, 103*(3), 581–583.

Nahmias, E. (2005). Agency, authorship, and illusion. *Consciousness and Cognition, 14*(4).

Neimeyer, G. J., & Rareshide, M. B. (1991). Personal memories and personal identity: The impact of ego identity development on autobiographical memory recall. *Journal of Personality and Social Psychology, 60*(4), 562–569.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology, 1*(1), 35–59.

Nietzsche, F. (1879/1986). Assorted opinions and maxims: Supplement to *Human, all too human*, tr. R. J. Hollingdale. New York: Cambridge University Press.

Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review, 111*(2), 486–511.

Nichols, S., & Stich, S. (2003). How to read your own mind: A cognitive theory of self-consciousness. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical essays* (pp. 157–200). Oxford: Oxford University Press.

Nichols, S., Stich, S., Leslie, A., & Klein, D. (1996). Varieties of off-line simulation. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind* (pp. 39–73). Cambridge: Cambridge University Press.

Oltmanns, T. F., Gleason, M. E. J., Klonsky, E. D., & Turkheimer, E. (2005). Meta-perception for pathological personality traits: Do we know when others think that we are difficult? *Consciousness and Cognition, 14*(4).

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Popper, K. R., & Eccles, J. C. (1977). *The self and its brain*. New York: Springer International.

Povinelli, D. J., & Prince, C. G. (1998). When self met other. In Michel D. Ferrari & Robert J. Sternberg (Eds.), *Self-awareness: Its nature and development* (pp. 37–107). New York: Guilford Press.

Schacter, D. L., Coyle, J. T., & Harvard Center for the Study of Mind Brain and Behavior (1995). *Memory distortion: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Schelling, T. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics, 59*(1), 1–22.

Thorne, A. (2000). Personal memory telling and personality development. *Personality and Social Psychology Review, 4*(1), 45–56.

Tooby, J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W. Kinzey (Ed.), *Primate models of hominid behavior*. New York: SUNY Press.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage, 14*(1), 170–181.
Wallace, A. R. (1889). *Darwinism: An exposition of the theory of natural selection, with some of its applications*. London: Macmillan.
Wegner, D. M. (2003). The illusion of conscious will. *Journal of Nervous and Mental Disease, 191*(2), 69–72.
Wegner, D. M. (2004). Précis of the illusion of conscious will. *Behavioral and Brain Sciences, 27*(5), 649–692.
Whiten, A. (Ed.). (1991). *Natural theories of mind: The evolution, development and simulation of everyday mind-reading*. Oxford: Blackwell.

Pascal Boyer
Department of Psychology and Anthropology, Washington University, St. Louis, MO, USA
E-mail address: pboyer@artsci.wustl.edu

Philip Robbins
Department of Philosophy, Washington University, St. Louis, MO, USA

Anthony I. Jack
Department of Neurology, Washington University, St. Louis, MO, USA

Received 18 August 2005
Available online 27 October 2005