Philosophical conceptions of the self: implications for cognitive science

Shaun Gallagher

Several recently developed philosophical approaches to the self promise to enhance the exchange of ideas between the philosophy of the mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self devoid of temporal extension, and the ‘narrative self’, which involves personal identity and continuity across time. The notion of a minimal self is first clarified by drawing a distinction between the sense of self-agency and the sense of self-ownership for actions.

This distinction is then explored within the neurological domain with specific reference to schizophrenia, in which the sense of self-agency may be disrupted. The convergence between the philosophical debate and empirical study is extended in a discussion of other cognitive sciences and that convey the breadth of best exchange of ideas between philosophy of mind and the other cognitive sciences. This review examines two important concepts of self: the ‘minimal self’, a self 

E.v.
The first approach involves various attempts to account for a ‘minimal’ sense of self. Even if all of the unessential features of self are stripped away, we still have an intuition that there is a basic, immediate, or primitive ‘something’ that we are willing to call a self. This approach leaves aside questions about the degree to which the self is extended beyond the short-term or ‘spacious present’ to include past thoughts and actions. Although continuity of identity over time is a major issue in the philosophical definition of personal identity, the concept of the minimal self is limited to that which is accessible to immediate self-consciousness. Certain aspects of the minimal self are relevant to current models in robotics. Furthermore, aspects of the minimal self that involve senses of ownership and agency in the context of both motor action and cognition can be clarified by neurocognitive models of schizophrenia that suggest the involvement of specific brain systems (including prefrontal cortex, supplementary motor area, and cerebellum) in the manifestation of neurological symptoms in this disorder.

A second approach to the concept of self involves conceiving of the self in terms of narrative. This notion was imported into the cognitive sciences by Dennett, but it might have a more complex significance than is indicated in Dennett’s account. The narrative self is extended in time to include memories of the past and intentions toward the future. It is what Neisser refers to as the extended self, and what Dennett calls a ‘nonminimal self’. Neuropsychological descriptions of episodic memory and its loss can help to circumscribe the neural substrates of the narrative self.

Self-reference and misidentification

These are a number of ways to understand the notion of a minimal sense of self. In this section, I approach the problem by discussing how we use the first-person pronoun in a self-referencing way that should never permit a mistake. This kind of self-reference has a feature that some philosophers call ‘immunity to error through misidentification relative to the first-person pronoun’. I will refer to this as the immunity principle (see Glossary). Once this principle is clarified we can ask whether, in actuality, it can ever fail and if so, what this might reveal. In the next section, I will explore this possibility in relation to a neurocognitive model of schizophrenia that requires us to make a distinction between two aspects of the minimal sense of self: the sense of self-ownership and the sense of self-agency.

Wittgenstein distinguished between two uses of the first-person pronoun in self-reference: ‘as subject’ and ‘as object’. Use of the first-person pronoun as subject might best be discerned by understanding what a speaker could be wrong about, and the kinds of questions that one could sensibly ask them. For example, if someone says ‘I think it is raining outside’, she could be wrong about the rain. It might not be raining. But it seems that she could not be wrong about the ‘I’. That is, she could not misidentify herself when she states that it is she who is thinking. So, according to Wittgenstein, the following question would be non-sensical: ‘Are you sure that you are the one who thinks it is raining?’. Such use of the first-person pronoun is immune to error through misidentification. By contrast, when we use the first-person pronoun ‘as object’ it is possible to misidentify ourselves. For example, in some experimental situations a subject’s arm may be deflected (that is, the subject is deprived of normal proprioceptive feedback about the position of their limb and therefore cannot keep track of it without vision). Their visual perception of arm movement is then manipulated through mirrors or videotape. In such cases, the subject might be led to say, ‘I am moving my arm to the left’, when in fact the basis for his judgment is a videotape of someone else moving their own arm to the left. In that case, the subject makes a mistake about who is moving their arm to the left. To say ‘I’ in such a case involves an objective misidentification of oneself.

Shoemaker suggests that the immunity principle applies only to the use of ‘I’ as subject because when we use the first-person pronoun as subject we are not actually attempting to identify ourselves. In other words, when I say ‘I’, I do not go through a cognitive process in which I try to match up first-person experience with some known criterion in order to judge the experience to be my own. My access to myself (my self) in first-person experience is immediate and non-observational; that is, it doesn’t involve a perceptual or reflective act of consciousness. In this sense, the immediate self that is referred to here is the pre-reflective point of origin for action, experience and thought. Are there any exceptions to the immunity principle? Is there any instance of someone using a first-person pronoun as subject, and being wrong in their reference? Following suggestions made by Frith and Frith about certain schizophrenic experiences (including auditory hallucination, thought insertion, and delusions of control in which subjects report that their body is under the control of other people or things), Campbell has proposed that such experiences might be counterexamples to the immunity principle. A schizophrenic patient who suffers thought insertion, for example, might claim that she is not the one who is thinking a particular thought, when in fact she is the one who is thinking the thought. The following example of a schizophrenic’s account of her own thought processes illustrates this: Thoughts are
comparator provides a sense of agency for movement; match at the feedback comparator generate two aspects of the ‘minimal self’ in normal experience (see text). Match at the forward comparator provides a sense of ownership for movement.

Fig. 1. The forward and feedback comparators. This model represents processes that generate two aspects of the ‘minimal self’ in normal experience (see text). Match at the forward comparator provides a sense of agency for movement; match at the feedback comparator provides a sense of ownership for movement.

A brief consideration of motor action will help to clarify two closely related aspects of minimal self-awareness: self-ownership – the sense that I am the author of the action. In the case of involuntary action, however, it is quite possible to distinguish between sense of agency and sense of ownership. I may acknowledge ownership of a movement – that is, I have a sense that I am the one who is moving or is being moved – and I can self-ascribe it as my movement, but I may not have a sense of causing or controlling the movement. That is to say, I have no sense of agency. The agent of the movement is the person who pushed me from behind, for example, or the physician who is manipulating my arm in a medical examination. Thus, my claim of ownership (my self-ascription that I am the one who is undergoing an experience) can be consistent with my lack of a sense of agency. Phenomena such as delusions of control, auditory hallucinations, and thought insertion appear to involve problems with the sense of agency rather than the sense of ownership.

It is thus the author of the action. In the case of involuntary action, however, it is quite possible to distinguish between sense of agency and sense of ownership. I may acknowledge ownership of a movement – that is, I have a sense that I am the one who is moving or is being moved – and I can self-ascribe it as my movement, but I may not have a sense of causing or controlling the movement. That is to say, I have no sense of agency. The agent of the movement is the person who pushed me from behind, for example, or the physician who is manipulating my arm in a medical examination. Thus, my claim of ownership (my self-ascription that I am the one who is undergoing an experience) can be consistent with my lack of a sense of agency. Phenomena such as delusions of control, auditory hallucinations, and thought insertion appear to involve problems with the sense of agency rather than the sense of ownership.

There is good evidence to suggest that the sense of ownership for motor action can be explained in terms of ecological self-awareness built into movement and perception10,16. By contrast, experimental research on normal subjects suggests that the sense of agency for action is based on that which precedes action and translates intention into action11,16. In addition, research that correlates initial awareness of action with scalp recordings of the lateralized readiness potential in motor cortex, and with transcranial magnetic stimulation of the supplementary motor area, strongly indicates that one’s initial awareness of a spontaneous voluntary action is tied to the anticipatory or pre-movement motor commands relating to relevant effectors17,18.

It turns out that some schizophrenic patients who suffer from thought insertion also make mistakes about the agency of various bodily movements. To explain this, Frith10 appeals to the notions of efference copy and comparator mechanisms that were originally used to explain motor control19,20. According to the most recent version of this model, a comparator mechanism operates as part of a non-conscious premotor or ‘forward model’ that compares efference copy of motor commands with motor intentions and allows for rapid, automatic error corrections21,22. Such a mechanism is consistent with the findings cited above. This comparator process anticipates the sensory feedback from movement and underpins an online sense of self-agency that complements the ecological sense of self-ownership based on actual sensory feedback17,23 (Fig. 1). If the forward model fails, or efference copy is not properly generated, sensory feedback may still produce a sense of ownership (‘I am moving’) but the sense of agency will be compromised (‘I am not causing the movement’), even if the actual movement matches the intended movement24.
lack of a sense of agency that is characteristic of these kinds of schizophrenic experience.

Following a suggestion made by Feinberg36, Frith posits a similar model for cognition – specifically, for thought and inner speech10. Phenomena such as thought insertion, hearing voices, or perceiving one’s own acts as alien, suggest that something is wrong with the self-monitoring mechanism. Frith’s model assumes not only that thinking, insofar as it is intended and self-generated, is a kind of action, but also that thinking has to match the subject’s intention for it to feel self-generated, as in the case of a motor action. This suggests that, although such intentions are not always consciously accessible, comparator processes that match intentions to the generation of thought and to the stream of thought below, respectively, a sense of agency and a sense of ownership for thought, in a similar way to motor action. If the mechanism that constitutes the forward aspect of this monitoring process fails, a thought occurs in the subject’s own stream of consciousness but does not seem to the subject to be self-generated or to be under the subject’s control. Rather, it appears to be an alien or inserted thought (Fig. 2).

Whether or not this kind of model is adequate to account for the phenomenon of inserted thought, I would suggest that the approaches taken by Frith and Campbell promise a way to explain in specific neurological terms the immediacy involved in the senses of self-ownership and self-agency, and in the immunity principle. Such aspects of the minimal self may find a neurological explanation in the proper workings of the mechanisms described above and are threatened by their failure.

The minimal self: embodied or disembodied?

Taking the immunity principle as a point of departure, there are two other directions that one could follow. The first explores the idea that there is an even more primitive and embodied sense of self than that involved in the use of the first-person pronoun. This approach pursues the implications of what developmental psychologists have recently discovered about the experience of neonates. The second involves a more abstract self-reflective access to first-person experience, which, among other things, leads to issues that concern AI applications in robotics.

Taking the first approach, are there any aspects of the minimal self that are more primitive than those identified in the immunity principle? In speaking about self-reference, we are already speaking of a self that is capable of linguistic communication – as the very least, the person is capable of using the first-person pronoun. If one considers that language and conceptual capacity develop in parallel, this might mean that the person’s immediate and pre-reflective access to the self already involves the mediation of a conceptual framework. Is it possible to speak of a non-conceptual access to the self – a more primitive self-consciousness that does not depend on the use of a first-person pronoun?

Bermúdez explores the many implications of this question37. Following on from Gibson’s ecological psychology, part of what Bermúdez calls non-conceptual first-person content4 consists of the self-specifying information attained in perceptual experience. When I perceive objects or movement in the external environment, I also gain information about myself – information that is pre-linguistic and non-conceptual. This is what Neisser calls the ecological self32. The fact that non-conceptual, ecological self-awareness exists from the very beginning of life can be demonstrated by the important role it plays in neonatal imitation. Neonates less than an hour old are capable of imitating the facial gestures of others in a way that rules out reflex or release mechanisms, and that involves a capacity to learn to match the presented gestures30,31. For this to be possible the infant must be able to do three things: (1) distinguish between self and non-self; (2) locate and use certain parts of its own body proprioceptively, without vision; and (3) recognize that the face it sees is of the same kind as its own face (the infant will not imitate non-human objects38). One possible interpretation of this finding is that these three capacities present in neonates constitute a primitive self-consciousness, and that the human infant is already equipped with a minimal self that is embodied, enactive and ecologically tuned39,40.

One can, however, move in a second direction and ask whether it is possible to capture and make explicit the pre-reflective minimal self in a reflective, and conceptually informed, introspection. In this case, one may well talk about the most abstract aspect of what we experience to be ourselves, even if it is mediated through reflection. Galen Strawson’s recent essays on the self make it clear that he is seeking the most basic and stripped-down version of a self that can still be called self41,42. He begins with a reflective description of his experience of the self. This phenomenologically reflective approach then naturally leads to a characterization of the self as a subject of experience. Thus, Strawson is led to define the self as a subject of experience that is a single (hirsus-free) mental thing. This is a momentary self without long-term continuity, and thus, without a history – ‘a bare locus of consciousness, void of personality’ (Ref. 3, p. 492).

On this view, a human being consists of a series of such transient selves, each one lasting only as long as a unique period of experience lasts, coming into existence and going out of

---

**Fig. 2. Neuror psychological model of inserted thought.** This model represents the loss of a sense of agency in some psychiatric illnesses such as schizophrenia. No match at the forward comparator deprives the system of a sense of agency for thought; match at the feedback comparator provides a sense of ownership for thought.
exist, without continuity. Despite the ‘local’ character of Strawson’s approach, that is, an approach that focuses on his own experience, the self that he seeks to define is not restricted to the human case. It would be entirely possible for the immediate self he describes to be instantiated in a non-human animal that has the right cognitive equipment. It has been argued that it might even be possible to create the minimal self in a machine, or more precisely, in a robot (see Box 1), but this would entail dropping the part of Strawson’s proposal that defines the self as a conscious subject of experience.

Furthermore, on Strawson’s view, it is not essential that this minimal self be embodied or active within an environment. The self-consciousness that captures this self is not ecologically embedded, but is one that operates on a conceptual level, already in possession of the concept ‘self’. Strawson is nevertheless a materialist, and considers the self as ‘mental thing’ to be a physical entity which, in the human case, is likely to be manifest in terms of brain processes. ‘What is metaphorically the case, however, is not always revealed in the phenomenological record: I can be conscious of myself as a minimal subject of experience that is a single mental thing, without being aware of the embedding or brain functions that make (or may make) the self consistent with self-reference that is immune from error through misidentification, as it specifies an access to the self that does not depend on applying empirical (in this case physical) criteria of identity. Even if in the case that the information that constitutes the minimal self is generated in ecologically embodied experience, and even if, in practice, a human being is capable of knowing that this is the case, one does not gain the self-consciousness that goes along with the minimal self by knowing this to be true or by being able to employ empirical criteria to verify it.

The self extended and mediated by narrative

So far we have considered only a minimal self, which is a concept of self that seems quite at odds with our commonsense conception of who we are. Surely we think and speak of ourselves as entities extended in time! Indeed, we think and speak of ourselves with individual identity, encompassing that continuous experience? What is the nature of this sense of a continuous self? Is it carried by a succession of momentary minimal selves that are tied together by real connections? Or
are momentary minimal selves simply abstractions from a more substantial continuity that is the more genuine self? The philosophical traditions are replete with a variety of answers to these questions.

One famous answer given by Hume suggests that the self consists of a bundle of momentary impressions that are strung together by the imagination. In effect, an extended self is simply a fiction, albeit a useful one because it lends a practical sense of continuity to life, but a fiction nonetheless. The narrative theory of self is a contemporary reading of this view. Dennett offers one version of this theory which he sees as consistent with recent developments in our understanding of how the brain functions. The consensus from contemporary neuroscience is that neurological processing is for the most part distributed across various brain regions, and it cannot be said that there is a real, neurological center of experience. Thus, there is no real simplicity of experience at one time nor real identity across time that we could label the self. At best, we might refer to a minimal biological self as something real. But the latter is nothing more than a principle of organization involving the distinction between self and non-self. Furthermore, this principle is found throughout living nature, and it is not something sufficient for the purpose of a coherent conception or identity over time, such as is found at the level of human experience. Humans, however, do have something more than this – we have language. And with language we begin to make our experience relatively coherent over extended time periods. We use words to tell stories, and in these stories we create what we call our selves. We extend our biological boundaries to encompass a life of meaningful experience.

Two things are to be noted from Dennett’s account. First, we cannot prevent ourselves from ‘inventing’ our selves. We are hardwired to become language users, and once we are caught up in the web of language and begin spinning our own stories, we are not totally in control of the product. As Dennett puts it, ‘for the most part we don’t spin them [the stories]; they spin us’ (Ref. 4, p. 418). Second, an important product of this spinning is the narrative self. The narrative self, however, is nothing substantially real. Rather, it is an empty abstraction. Specifically, Dennett defines a self as an abstract ‘center of narrative gravity,’ and likens it to the theoretical fiction of the center of gravity of any physical object. In the case of narrative gravity, however, an individual self consists of the abstract and movable point where the various stories (of fiction or biography) that the individual tells about himself, or are told about him, meet up (Fig. 3a).

The notion of a narrative self-constitution finds confirmation in psychology and neuroscience. In the former, Neisser’s concepts of the extended and the conceptual self, initially explained in terms of memory, have been enhanced by considerations of the role that language and narrative play in developing our own self-concept. In the realm of neuroscience, Gazzaniga has suggested that one function of the left hemisphere of the brain is to generate narratives, using what he terms an ‘interpreter’. Gazzaniga proposed this interpreting function based on his studies of split-brain patients. In these patients, the left hemisphere has no internal access to right-hemisphere experience because the corpus callosum has been severed. Nonetheless, in precisely designed experimental circumstances, the left hemisphere devises interpretations for meanings, actions and emotions produced by the right hemisphere. Such interpretations show consistency with the experiential context belonging to the left hemisphere rather than with the original right-hemisphere context. The left hemisphere, for example, might remain ignorant of the content or cause of an emotion generated in the right hemisphere, but the left-hemisphere experience of that emotion motivates an interpretation of the event in terms relevant to the content available to the left hemisphere. In Gazzaniga’s model, the interpreter weaves together autobiographical fact and inventive fiction to produce a personal narrative that enables the sense of a continuous self. Gazzaniga, however, contends that the self, in this regard, is not a fiction because the normal functioning of the interpreter tries to make sense of what actually happens to the person. At most, in the non-pathological case, it may be only ‘a bit fictional’ (Ref. 41, Trends in Cognitive Sciences – Vol. 4, No. 1, January 2000).
Outstanding questions

- What relationship exists between the minimal self and the narrative self? Is one generated from the other? Do they operate independently of each other?
- Shenhavan7 has maintained that immunity to error does not necessarily extend to episodic memory. What status do truth-claims concerning episodic memory have?
- Because the sense of self-agency is absent in the case of involuntary others, what mechanisms at the psychological level integrate or adjudicate these diverse constitutions?
- In narrative theory, self-constitution is meant to imply a situation in which the self is both the narrator and the narrated. If there is more than one narrative interpreter?
- What is the precise neurological mechanisms involved in the left-hemisphere interpreter?
- In narrative theory, self-constitution is meant to imply a situation in which the self is both the narrator and the narrated. If there is more than one narrative interpreter (if we are also constituted by stories about ourselves told by others), what mechanisms at the psychological level integrate or adjudicate these diverse constitutions?

Concluding remarks

In a recent book, Damasio has insightfully captured the difficulty involved in expressing the interrelations between the minimal (‘core’) self and the narrative (‘autobiographical’) self42. The difficulty is due to complexities that are apparent on both the personal and the sub-personal, neurological levels. Episodic memory, which is necessary for the construction of the narrative self, is subject to constant remodeling under the influence of factors that include innate and acquired dispositions as well as social and cultural environments. The registration of episodic memory as ‘my’ memory of ‘myself’ clearly depends on a minimal but consistently reiterated sense of self that I recognize, without error, as myself. In some respects, as Damasio insists, this depends on narrowly defined, embodied capabilities and feelings. In other regards, however, the core features of the self are constantly being reinterpreted by the narrative process. In the neurological terms that Damasio uses this means that there are extremely complex demands made on the processes that link early sensory cortex that hold information on the minimal or core self, and convergence or dispositional zones that contribute to the generation of the narrative self. In this regard, he makes it clear that at present the neuroscientist, like the philosopher, can offer, at best, informed speculation on these processes.

In this review, I have tried to show that philosophical ideas about the self can be aligned with, and can inform, current ideas in cognitive science. I also believe that philosophers can learn about the nature of the self from psychologists, neuroscientists and other cognitive scientists. Thus, collaborative efforts between philosophers and scientists promise to open up more subtle and sophisticated avenues of research, which will define more fully the concept of the self.

Acknowledgements

I thank S-I. Bakken, G. Eckstein, J. Neisser, and anonymous reviewers for helpful comments on an earlier version of this paper. Parts of my research were supported by a Fellowship at the NIH Summer Institute on Moral, Self, and Psychopathology, directed by J. Whitby and L. Saxe at Cornell University in 1996, and by a sabbatical leave from Canisius College in 1998.

...
References
11. Campbell, J. Schizophrenia, the space of reasons and thinking as a motor process. The Abolit (in press)

Trends in Cognitive Sciences online – tics.trends.com

The online version of Trends in Cognitive Sciences contains a simple search and browse facility that will enable you to view full-text versions of all articles published in the journal since January 1998. Browsing and searching the articles is straightforward and designed around the need for simple navigation. Articles can be viewed in three display formats: text-only, text with pop-up figures, or as PDF files for high-quality printing.

Personal print subscribers benefit from online access to Trends in Cognitive Sciences online by accessing the site using a unique subscription key. For information about how to obtain your key please visit http://www.trends.com. The subscription key is important and should be kept safe as it is your unique access code to your online journal for the remainder of the year. Institutional access is currently available through Elsevier’s ScienceDirect (http://www.sciencedirect.com) services.

Subscribers and non-subscribers alike will continue to have free access to full content lists, abstracts, search facilities and the e-mail contents alerting service. It is also possible to download full-text articles on a pay-per-view basis.

The following recent articles are just some of those available in the fully searchable archive: