Reconceiving the Self

I. Introduction

The concept of the self is notoriously slippery.[1] The "problem of the self" is often construed as a metaphysical problem, analogous to the puzzle regarding the identity of Theseus' ship. Here, the basic issue is how a person can remain the same while undergoing change. A very different set of concerns is found in the literature of how one ought to live. In this literature, the focus tends to be on the moral and practical decision making practices and how these practices reflect on or constitute the self. Both literatures are dominated by familiar methods of philosophical investigation, including thoughtful reflection on our own practices, thought experiments to test our intuitions, and introspection of the phenomenology of our experience.

There is also a growing empirically based literature in which there is no general agreement on a single problem of the self; instead, philosophers follow the empirical data to carve out niches where interesting questions involving some concept of self arise. For example, in comparative ethology there are now a number of animals who appear to pass the mirror self-recognition task and philosophers theorize about what passing such a task implies.[2] Or, for another example, the literature on folk psychology and theory of mind investigates the relationship between first-person and third-person mental concepts.[3]

The empirical research from the last two decades undermines some of the core assumptions of the more traditional philosophical conceptions of the self and, perhaps more significantly, should shake our faith in traditional philosophical methods.[4] In this paper, I focus on the work of psychologists and neuroscientists which suggests that much of our behavior, feelings, and judgments are driven by unconscious mental states. The fact that these unconscious mental states are largely inaccessible to first person introspection should make us wary of relying solely on the first-person perspective when it comes to understanding the self. I begin with a very brief discussion of four features that characterize both the philosophical and the common conception of the self.

[1] The fragmentation of the literature has spurred at least one philosopher to argue for abandoning the concept. See Eric Olson (1999). For a discussion of the various uses of "self" by psychologists, see Mark Leary and June Price Tangney (2003).

[2] Researchers inconspicuously place a mark on the animals, which have become accustomed to mirrors. If the animal displays curiosity about the mark (by rubbing, scratching, peering etc.) then it passes the mirror recognition test. See Gallup (1998) and Povinelli (1998) for differing interpretations of what passing this test implies about animals' self concept.

[3] See Davies and Stone (1995) for both philosophical and psychological perspectives on this topic.

[4] Eric Schwitzgebel has written convincingly about the unreliability of introspection. See especially his (2008). See John Doris (2009) for arguments against traditional conceptions of personhood. For a classic paper undermining our confidence in first person reports see Nisbett and Wilson (1977).

I begin with a brief discussion of our idea of the self, review some of the empirical literature, and end with a brief discussion of three models supported by the empirical work.

II. Four features of our concept of self

There are at least four characteristics of our concept of the self that are crucial to both the ordinary notion and the philosophical concerns. To begin, as I mentioned above, there is the fact that selves explain the unity of a human person. Selves explain how people are connected to their past and future. Bodies change, mental states change, but people tend to believe that in some sense of the word they are the *same* people now that they were in the past and will be in the future. The connection between past, present and future selves is sometimes called diachronic unity. A different sort of unity – synchronic unity – can also be explained by appeal to selves. This is the sense we have that all the various experiences that occur at a time are occurring to the same subject. For example, the haptic experience I have of the keyboard, the sound of the cardinal singing outside, the sight of the dog lying on the carpet are all happening now to the same person, myself. The various experiences are unified in a single subject.

A second feature of selves which is important to both philosophers and the general public is the notion of agency. Selves are agents who initiate and direct their actions in the world. To attribute a self to a stick or a rock is to make a gross category mistake – though one can make sense of the notion of attributing selves to living creatures who are less complex than humans. Agency makes us responsible, both ethically and practically, for our actions. It's this aspect of the self which has been the main concern of ethicists and philosophers of action.

Personality or character traits are also part of the ordinary conception of selves; this aspect has been of less interest to philosophers, but psychologists have frequently assumed that personality traits are a central part of our concept of self. People assign themselves certain traits and not others, and more importantly they identify with these traits. Personality or character traits also play a role in distinguishing one self from another.

Finally, there is the fact that we are self-aware. Arguably, this feature sets human selves apart from other animals. It's conceptually possible that other animals have selves that share the first three features. My dog, for example, seems to have a unified experience, is a practically responsible agent, and has distinctive personality traits, but like most, if not all other animals, my dog lacks the kind of self-awareness that characterizes human mental life. Our ability to be aware of our mental states and actions, both past and present, makes possible for us, in a way that it isn't for other animals, to form a concept of the self.

These are not isolated features; they are related to each other in various ways. For example, the common American belief in the constancy of character is part of the story behind the sense that selves are diachronically unified. And the fact that we can think about ourselves and others *as selves* is part of why we are capable of moral agency. I'll argue in the next section that all of these core features are undermined by the empirical literature. There is good reason to believe that we are less unified, less in control, more instable, and less aware.

III. The empirical literature

The literature on what's sometimes called "automaticity" or the "new unconscious" is vast and growing daily.[5] While the idea that we sometimes act unconsciously or automatically is not new, the research indicates that it is more widespread and more sophisticated than we might have supposed. I begin with discussion of an oft-cited experiment by the psychologist John Bargh and his colleagues. The experiment began with a language task in which the participants were exposed to either words related to rudeness or politeness or neither (the control). The participants had been told in advance that the experiment would involve two distinct tasks. After they had completed the language task the subjects were given an opportunity to interrupt a conversation between the experimenter and a confederate in order to ask about when the next task would begin. Whether or not the participant would choose to interrupt the conversation was strongly predicted by which group the participant was in. Those who had been exposed to the "rude" words interrupted 67% of the time, while those who had been exposed to the "polite" words interrupted only 16% of the time. The control group fell in the middle, interrupting 32% of the time (1999, 466). The experiment has been repeated many times with variation. One of the most fascinating versions exposed participants during the language task to elderly stereotypes (*wrinkle, Florida*) and then measured how slowly they walked down the hall to exit the building. Those who had been exposed to the elderly stereotypes walked significantly more slowly down the hall (1999, 466).

Even more intriguing, I think, is that the participants are unaware that they are acting rudely/walking slowly and, when their behavior is pointed out, they are unable to provide the reason for their actions. In these cases, we have examples of external stimuli causing unconscious mental states which in turn cause certain types of behavior and the subject of these mental states seems to have no awareness of

[5] See *The New Unconscious* (2005) ed. Hassin, Uleman, and Bargh for a comprehensive overview.

the external stimuli, their mental states, or even their behavior. (Subjects, of course, realize that they are walking, but they don't realize that they are walking slowly.)

A more dramatic demonstration of the lack of connection between conscious and unconscious behavior can be found in the work of Pierre Fourneret and Marc Jeannerod (1998). Fourneret and Jeannerod asked subjects to use a stylus on a graphic tablet to trace a line represented on a computer screen. The participants' hands were blocked from view so they received no visual feedback, but their hands' positions were graphically represented on the screen. This being a psychology experiment, there was, of course, a trick. The experimenters had preprogrammed the computer so that the representation of what their hand was doing would move in a different direction (the bias ranged from 2-10 degrees from the straight line the subject was supposed to be tracing) from the actual direction the participants' hands were moving. Surprisingly, however, none of the participants noticed. Here we have people completely wrong about what their own hands are doing, without any hint that they might be wrong.[6] Fourneret and Jeannerod themselves conclude, "normal subjects appear to be poorly, if at all, aware of the details of their motor performance and to be unable to correctly monitor, consciously, the signals generated by their own movements" (1137).

The literature on automaticity has been helped by the development of a new research method, the implicit association test (IAT). To take the test, a subject sits in front of a computer with her index finger of each hand over the 'e' or 'i' keys. The task is to sort things into categories. So, for example, a subject will begin by sorting close-up pictures of white and black faces into the categories European American and African American. She will then sort words into two categories of 'good' and 'bad'. Things get more complicated with the final stage. The subject is ordered to pick out black faces and positive words with the 'e' key and white faces and negative words with the 'i' key. Then the categories are switched, with white faces and positive words being categorized together and black faces and negative words being categorized together. The task is timed; the subject must respond quickly to the item on the screen, or the response is thrown out. The idea behind the test is there are implicit associations that hold among our concepts and this test reveals them. For example, if a subject is quicker to respond to the combination [Black and Positive/White and Negative] then the test indicates that she has a stronger association between her Black concept and positive concepts than between her White concept and positive concepts (Greenwald, et al.,

[6] Perhaps this shouldn't be such a surprise. Many people report having had the experience of thinking that they are playing a video game only to realize that the machine was running a simulation and their frantic actions with the joystick were completely impotent.

2002, 18). In a meta-analysis, Greenwald and colleagues found that the IAT is a better predictor than self-reports for certain topics like Black-White interracial behavior and intergroup behavior (2002, 28). On other words, a subject's consciously held beliefs are less useful a predictor of behavior than her unconscious implicit beliefs; moreover, these often will be in conflict.

This is a very small sample of a large and growing literature that undermines some of our most cherished beliefs about ourselves. All four of the characteristics discussed in the previous section – unity, agency, personality, and self-awareness – are implicated in these studies. Much of our behavior is driven by unconscious processes that are both inaccessible to conscious introspection and often in conflict with our conscious desires. We are less unified, less in control, less stable in personality, and often plain wrong in our assessments of our selves.

IV. What now?

Two models of the self immediately suggest themselves in response to the empirical literature. According to the *multiple selves* model, the human mind possesses multiple processes that end in action and are fairly isolated from one another. This is well illustrated by the Titchener illusion. When people with normal vision look at the Titchener illusion it appears that the middle circle surrounded by little circles is bigger than the middle circle surrounded by bigger circles. This being an illusion, the middle circles are the same size. And some part of the human mind knows this. When we reach out to grab the inner circles our fingers form a grasping position which is identical in diameter in both cases. Some part of our minds is not fooled by the illusion and it is this part which is in control of our grasping behavior. Essentially, the unconscious overrides our conscious visual experience and directs our behavior.[7] The neuroscientist V. S. Ramachandran uses this example, among others, to argue that our minds consist of lots of "zombie selves" that direct and control our behavior yet remain inaccessible to us, at least from the first-person perspective.[8]

Endorsing the multiple selves model is to give up on the traditional philosophical pursuit of an account of personal identity. According to this model, the self is merely an illusion. Our concept of ourselves might be as a single entity in control of our lives, but this concept is empty; there is no single entity that makes decisions, has experiences, and controls our behavior. This is fundamentally a skeptical

[7] This example is discussed in Ramachandran (1998).

[8] A more detailed account of the multiple selves model can be found in Humphrey and Dennett (1989).

position. I agree that the empirical evidence forces us to revise traditional conceptions of the self, but I'm not (yet) convinced that we need to give up on the project entirely.

A very different suggestion is what I call the *new dualism* model. Underlying this model is a commitment to dual process theory, the idea that we have two basic kinds of processes, conscious and automatic. On a new dualist model of the self, the real self gets identified with the conscious processes and the automatic processes are assumed to be not essential or significant when it comes to issues about the self. You see something like this in Frankfurt style accounts of free agency – only those acts which are accompanied by a second order volition count as free, the rest are merely the acts of a wanton or animal. Frankfurt's account is endorsed explicitly by the psychologist Keith Stanovich in his book *The Robot's Rebellion*. In this book, Stanovich uses Richard Dawkins' distinction between the genes and the vehicle of the genes to argue that we, as vehicles, are in constant battle against our own genes. The genes have one desire – to copy themselves –while the vehicles (us) have entirely different desires.[9] He argues that the genes' desire to replicate drives much of the unconscious behavior and that what distinguishes us as persons is our rational capacities. Using our rational abilities is how we thwart the genes – we humans are the robots rebelling in his book title. And, Stanovich sometimes seems to claim, it is with our rational self that our true self lies.[10]

The new dualism model gives up on the unity of the self in a way very familiar to philosophers. From Plato's arresting image of the soul as a charioteer to Descartes' elevation of those mental states not infected by the senses, there has been an influential philosophical tradition of imagining conflict within the self being subdued by reason. The implication, sometimes explicit and sometimes implicit, is that the "true self" is identified with the rational element. Like many others, I'm not persuaded by such a view of human nature. In part, this is because I reject the notion that the rational element is any more essential or "true" than other features of human life; this view severely underestimates humans' complicated social emotions for example. More important for the purpose of this paper, the new dualist model gives up on the project of unification, by rejecting the unconscious mental states as part of who we are, rather than looking for an inclusive conception of the self.

My own view is that we should look for an inclusive conception of the self and this will require abandoning the commonplace view that our selves are constituted primarily by conscious mental states. One way to accomplish this shift in perspective is to take much more seriously the fact that we are creatures

[9] Stanovich recognizes, as do I, that the use of 'desire' here is metaphorical.

[10] Stanovich's view is more complicated than this brief sketch allows. In particular, he does allow that there will be times when the automatic behavior gets it right.

with an evolutionary past.[11] Taking our evolutionary past as a starting point for our accounts of the self is helpful for many reasons. First, it acts as a counterpoint to the constant pull towards dualism. Thinking of ourselves in a dualistic way comes easily for many of us and we need help in resisting it.[12] Second, it reinforces the essentially social nature of humans, which informs how we think about what kinds of selves we are. Much of the Western tradition has viewed the self in an atomistic, individualistic way. Along with many other philosophers, I think that this has been a mistake. Recent empirical work in the social sciences has begun to emphasize how surprisingly unique humans are in their understanding and ability to cooperate with others.[13] Philosophers need to incorporate these findings into our revised conception of the self. Third, making the evolutionary history of humans more central to our understanding of our conception of ourselves pushes us to think more about human behavior and less about human mental life that does not result in behavior.

One of the benefits of this approach is that this revised conception appears to lead to more accurate self-knowledge. Recent work by Emily Pronin, Jonah Berger and Sarah Malouki (2007) suggests that we would be well served to start paying more attention to our behavior and less attention to our mental states. In five studies exploring subjects' conformity judgments the researchers found a persistent bias when it came to first person attributions of conformity compared to third person attributions of conformity. That is, subjects considered themselves as less likely than their peers to conform across a range of situations. More interesting for my purposes, Pronin et al also investigated the source of this bias. They found evidence across all five studies that the subjects' fell prey to what the researchers call "the introspective illusion." In brief, the introspective illusions occur when subjects pay more attention to their own mental states and not enough attention to their behavior in making self-ascriptions. This is exacerbated by subjects' belief that introspective information about their own mental states is more valuable than introspective information about others' mental states. The upshot: more accurate self-assessments are made when a subject discounts her mental states and focuses on her behavior.

In conclusion, I hope to have shown that there is empirical literature relevant to philosophers' theorizing about the self, that this literature has revisionary implications for certain conceptions of the self, and also undermines many of the first-person methods that philosophers have relied upon in drawing their accounts.

[11] Here I am in agreement with Stanovich, whose account is thoroughly informed by the fact that we are biological organisms shaped by evolutionary forces.

[12] See Paul Bloom, *Descartes Baby* for an argument that dualistic thinking comes early and naturally to humans across the world.

[13] See Sarah Blaffer Hrdy, Michael Tomasello, Frans de Waal, Elliot Sober and David Sloan Wilson for empirical work that focuses on the significance and uniqueness of human social abilities.

References

Balcetis, Emily Dunning, David and Miller, Richard (2008). "Do Collectivists Know Themselves Better than Individualists?" *Journal of Personality and Social Psychology*, 95:6, 1252-1267.

Bargh, John and Chartrand, Tanya (1999). "The Unbearable Automaticity of Being." *American Psychologist,* 54:7, pp. 462-479.

Davies, Martin and Stone, Tony. (eds.) (1995) *Folk Psychology: The Theory of Mind Debate*, Oxford: Blackwell Publishers.

Doris, John. (2009). "Skepticism about Persons." Philosophical Issues 19: Metaethics, 57-91.

Dutton, Donald G. and Aron, Arthur P. (1974). "Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety." *Journal of Personality and Social Psychology*, 30:4, 510-517.

English, Tammy and Chen, Serena (2007). "Culture and Self-Concept Stability" *Journal of Personality and Social Psychology*, 93:3, 478-490.

Fourneret, Pierre and Jeannerod, Marc (1998). "Limited conscious monitoring of motor performance in normal subjects." *Neuropsychologia*, 36:11, 1133-1140.

Frankfurt, Harry G. (1971). "Freedom of the Will and the Concept of a Person." Reprinted in *The Importance of What We Care About* (1998) New York: Cambridge University Press, 11-25.

Gallagher, Shaun and Shear, Jonathan (eds.) (1999). *Models of the Self.* Exeter: Imprint Academic.

Gallup, Gordon G., Jr. (1998) "Can Animals Empathize? Yes," *Scientific American Presents*(Winter 1998), 66-71.

Gendler, Tamar Szabo (1999). "Exceptional Persons: On the Limits of Imaginary Cases." In Gallagher, Shaun and Shear, Jonathan (eds.) (1999). *Models of the Self.* Exeter: Imprint Academic.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (2009). "Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity." *Journal of Personality and Social Psychology*, 97, 17-4

Humphrey, Nicholas and Dennett, Daniel C. (1989) "Speaking for Our Selves: an Assessment of Multiple Personality Disorder." *Raritan*, 9:1, 68-98.

Olsen, Eric (1999). "There is No Problem of the Self." In Gallagher, Shaun and Shear, Jonathan (eds.) (1999). *Models of the Self.* Exeter: Imprint Academic.

Leary, Mark R. and Tangney, June Price (2003). "The Self as an Organizing Construct in the Behavioral and Social Sciences." In Leary, Mark and Tangney, June Price (eds.) (2003) *Handbook of Self and Identity*, New York: Guilford Press.

Hassin, Ran, Uleman, James and Bargh, John, eds. (2005) *The New Unconscious.* New York: OUP.

Heine, Steven J. (2001). "Self as Cultural Product," *Journal of Personality*. 69:6, 881-906

Povinelli, Daniel J. (1998) "Can Animals Empathize? Maybe not," *Scientific American Presents* (Winter 1998), 67-75.

Pronin, Emily, Berger, Jonah, and Malouki, Sarah (2007). "Alone in a Crowd of Sheep: Asymmetric Perceptions of Conformity and Their Roots in an Introspection Illusion." *Journal of Personality and Social Psychology*, 92:4, 585-595.

Ramachandran, V.S. and Blakeslee, Sandra (1998). *Phantoms in the Brain*. New York: Quill.

Schwitzgebel, Eric. (2008). "The Unreliability of Naive Introspection." *Philosophical Review,* 117, 245-273.

Stanovich, Keith E. (2004). *The Robot's Rebellion*. Chicago: University of Chicago Press.

Strawson, Galen, ed. (2005). *The Self?* Malden, Mass: Blackwell.

Wilson, Timothy (2002) *Strangers to Ourselves*. Cambridge, Mass: Belknap Press.