PREPRINT: Gallagher, S. 2011. The self in the Cartesian brain. Perspectives on the Self: Conversations on Identity and Consciousness. *Annals of the New York Academy of Sciences* 1234: 100–103.

The self in the Cartesian brain

Shaun Gallagher
Moss Professor of Excellence in Philosophy
University of Memphis (USA)
Research Professor of Philosophy and Cognitive Science
Universit of Hertfordshire (UK)

It seems odd that after 370 years or so, since Descartes published his *Meditations* (1641), we are still wrestling with his thought. Not just philosophers, but scientists who study the mind, as well (see, e.g., Edelman 2006). By the time Descartes himself arrives at his Sixth Meditation, he is wresting with his own thought. Insofar as he defined the self as a thinking thing (*res cogitans*) in the Second Meditation, he seems to be debating with himself as he attempts to work out how precisely the body comes into play with the mind. That the two things (mind and body) interact, he is sure. His theoretical dualism is still a problem, but practically speaking he knows that body and brain (at least the pineal gland) have something to do with cognition and self. And he is almost certain of this. The degree of certitude he has about bodily involvement in cognition, however, is not as great as the certitude he found in the Second Meditation concerning the self as thinking thing. This seemed to him to be the one thing beyond doubt. He gains a high degree of certitude about the body, however, only by bringing God into the picture.

The very quick version of what happens between the Second and the Sixth Meditation is that Descartes attempts to prove (1) that God exists, (2) that (by definition) God is not a deceiver, and (3) that since our capacities for judgment are God-given, we should trust our judgments as long as we are careful to follow proper methods. Hence our judgment that the body has something to do with the self and our cognitive life seems clear. The problem is, whether God exists or not, Descartes argument for the existence of God is questionable, and we have good reasons to doubt its validity. Of course the very hypothesis that God exists is not something with which modern science wrestles. We've learned that this is simply a question that empirical science cannot address. As a result, absent Descartes' theological solution to his epistemological quandary, science and philosophy are thrown back to the earlier Meditations, and specifically, setting God aside, they find themselves wrestling with the evil demon.

The evil demon is a thought experiment devised by Descartes in the First Meditation to buoy up the extremely strong doubt that gets his meditations off the ground. There seems to be no way to know that our entire life experience and all of our thoughts have not been a large and complex illusion caused by an all-powerful demon (or if you are a fan of the *Matrix* films, an all encompassing matrix). Descartes is not saying that this *is* the case; only that there is no way to tell that it isn't. Hence, we should doubt everything. Except for the one thing that we cannot doubt, since even if the evil demon fools us into thinking that we are thinking, we are still thinking. *Cogito ergo sum*.

For some philosophers and scientists who think about the self, this is more or less where we still are. In place of the evil demon, however, we have the operations of our brains, which are seemingly something very real, but which, like the evil demon generate an illusion, or a set of illusions. Free will, for example, is one such illusion, as Daniel Wegner (2003) argues, in part on the basis of the Libet experiments that purportedly show that brain activity (the readiness potential) correlated with a particular action predates our conscious decision about performing that action by at least several hundred milliseconds (Libet 1985; Libet et al. 1983). Although we think that we are free, and although we have a sense of agency for our actions, this is an illusion perpetrated by the brain, perhaps to make our lives more interesting. It may also be the case that the world as we perceive it is a grand illusion, since, of course, that experience is sketchy at best and the product of certain neuronal representations (this is an argument that Noë [2002] pieces together, attributes to e.g., Blackmore et al. [1995], and then criticizes). Mess about with those neuronal processes and the world starts to look different, as it might to a person who is schizophrenic. Such internalist, neuro-centric views pull the certitude that Descartes had found in the Second Meditation back into the doubt that reaches a crescendo at the end of the First. If we think certainly that we are thinking things, it turns out that we are "nothing but a pack of neurons," as Francis Crick (1995) so nicely put it.

This is one influential position in regard to what we call "the self." Thomas Metzinger, for example, defends this view: "no such things as selves exist in the world: Nobody ever *was* or *had* a self" (2003, 1). The self, if not an illusion, is something close to it; what we call the self (and seemingly experience as the self) is nothing but a "self-model" generated by the brain. The self that Metzinger denies is precisely the Cartesian self, the thinking *thing*, which, as construed by Descartes, is a thinking substance. There is nothing substantial about the self.

Galen Strawson (1999a) draws a different model of the self. On his view the self is nothing more than a momentary experience (perhaps no more than 3 seconds in duration), which he defines as a distinct, mental, single, subject of experience. Its average life-span may be around 3 seconds because there is neurological evidence that the brain generates a coherent window of experience that is approximately that duration (see e.g., Pöppel 1978; Strawson 1999a, pp. 9-10).

In a certain way, there is something very Cartesian about these self-models. Not just in the sense that they are Second-Meditation "things," or First-Meditation matrix-like illusions, but because in a very real sense, and despite their residence in the brain, they are disembodied. Of course these selves may seem to be embodied, and we certainly experience our-selves as such, but if we start to look closer we find ourselves wrestling with the puzzles of the Sixth-Meditation, this time without God, although we still have our brains. And that's were we find our bodies. The prominent view in neuroscience and neurophilosophy is not that the brain is in the body, but that the body is in the brain, reducible in all important aspects to what Melzack (1990) calls a *neuromatrix*. The idea that *the body is in the brain* is not simply another thought experiment dreamt up by philosophers, although there is such a thought experiment called "the brain in the vat" (Putnam 1992); no, it's also an idea to be found in the most recent neuroscience (see Berlucchi and Aglioti 1997; 2010; Dolan 2006; Giummarra et al. 2008; Graziano & Botvinik 2001; Jackson et al., forthcoming). As Antonio and Hannah Damasio put it, "we (mentally speaking) exist in our bodies, and … our bodies exist in our minds" (2006,

15), and as they go on to show, this means that "the construction of the self would simply not be possible if the brain did not have available a dynamic representation of its body" (p. 21). Of course the claim that the body is in the brain is more rhetorical than metaphysical; no one claims that the physical body is literally in the brain, and there is plenty of evidence available to show that the body regulates the brain as much as the brain regulates the body. Neuroscience is rightly focused on understanding the brain processes that are involved in such two-way regulations. Still, the rhetoric sometimes leads the science and the philosophy when it comes to thinking about precisely what the nature of the embodied self is. The brain-in-the-vat thought experiment attempts to show that in principle (even if not in reality) to whatever extent the self is embodied, it is so only because the brain generates the representation in this way.[1]

In seeming opposition to the construal of the self as nothing more than a product of brain processes, some theorists have turned to narrative theory. On this view, "selves are inherently narrative entities" (Schechtman 2011, p. 395), where a narrative self, in contrast to the neural or minimal self, is defined as "a more or less coherent self (or self-image) constituted with a past and a future in the various stories that we and others tell about ourselves" (Gallagher 2000, 15).

There are various theories about the nature of the narrative self, but there is one that leads us right back to neuroscience. Daniel Dennett, for example, defines the self as a "center of narrative gravity" – an abstract and non-real point of intersection where the various stories about oneself come together (1991, 418). Although this doesn't make the self an illusion, Dennett takes an ambiguous position on precisely what the narrative aspect of self is. On the one hand, narrative is the product of linguistic processes that are generated in the brain, get projected into the world, and loop back into the brain. On the other hand, Dennett sometimes describes the narratives as lines of sub-personal brain processes that compete with each other to rise to the level of consciousness. Michael Gazzaniga (1998) follows this line more closely and suggests that narratives are generated in the neural processes of a left-hemisphere interpreter which in many cases simply confabulates our story in order to make sense of experience. This comes closer to the idea that our narrative self is something of an illusion generated by the brain. Our self-model is a narrative model that is ultimately cashed out in brain processes.

The brain processes involved in generating a sense of self can be very complex depending on what one means by the self. In a recent review article, Gillihan and Farah (2005) show that even when studies focus on a specific self-related representation, such as self-face recognition (judged in contrast with another person's face, or morphed faces), different methodologies and different subject groups will identify different areas of the brain for this function, including right hemisphere, left hemisphere, left anterior insula, putamen, and pulvinar, the right anterior cingulate cortex, and globus pallidus, left fusiform gyrus, anterior cingulate cortex, right supramarginal gyrus, superior parietal lobule, and precuneus, right middle, superior, and inferior frontal gyri, and right insula, hippocampal formation, and lenticular and subthalamic nuclei, the left prefrontal cortex (inferior and middle frontal gyri), right middle temporal gyrus, left cerebellum, as well as

[1] "A bodiless brain in a vat could certainly enjoy the *phenomenal* experience of holding a paper like this one in its *own* hands right now. The phenomenal content of your bodily self-representation is entirely determined by internal properties of your brain" (Metzinger 2005, 3.3.2; see Gallagher 2005 for discussion).

parietal lobe and lingual gyrus. Indeed, it looks like the entire cortex is specialized for self-referential processing. Gillihan and Farah, however, stuggest that studies of self-trait descriptions (what they define as a psychological component of self) provided no clear results for specialized brain areas because of various confounds. "The different ways in which the self–nonself distinction is confounded with other distinctions across studies are likely to account for the different patterns of activation in different studies … even when the same aspect of the self is under study" (Gillihan and Farah 2005, 94). More generally, however, Legrand and Ruby (2009) have shown that the diverse areas implicated in self-referential experience are in fact not areas of activation exclusively for self. The various brain areas frequently identified as self-specific brain are activated for cognitive processes that apply not just to self, but to other persons, and even to objects. It seems that the self is everywhere in the brain, or it's nowhere in the brain. This, however, entirely depends on how one defines the self (Vogeley and Gallagher 2011).

There are two points to make in regard to the neuroscience of the self. The first is that within neuroscience itself there is no clear picture of what one means by 'self'. If, as the recent reviews seem to suggest, the self is everywhere and nowhere in the brain, it's only because there is no clear consensus operating in neuroscience about what *self* means. Is the self an illusion; is it an embodied reality; is it a narrative entity? These are only three possibilities. William James (1890) defined 4 different conceptions of the self; Ulrich Neisser (1988) defined 5; Galen Strawson (1999b) identified 26 conceptions. It's not clear that this multiplication of selves is progress, but it is clear that the self is a complex phenomenon. It is also important to note that at least on some conceptions, the self is not something that depends solely (or solipsistically) on mental or brain processes that belong to the singular individual. If the self is social or intersubjectively constituted, as James, Neisser, and Strawson all suggest, then it has to be more than something that can be explained in terms of neuronal processes located in individual heads.

This leads to my second point. No one of these disciplines, whether neuroscience, philosophy, psychology, or any other, can claim to provide a full account of what seems to be a multi-dimensional and highly complex phenomenon. It is only a partial story to say that the brain is involved in the origins of the self. While there is no denying that the brain plays an important role, there is also no denying that to understand the self, like so many other complex phenomena, like consciousness, space, time, embodiment, or our relations with others, one requires many different arts and sciences.

References

Berlucchi G, Aglioti S 1997. The body in the brain: neural bases of corporeal awareness. *Trends in Neuroscience* 20: 560–564.

Berlucchi G, Aglioti S 2010. The body in the brain revisited. *Experimental Brain Research* 200: 25–35.

Blackmore, S.J., Brelstaff, G., Nelson, K., Troscianko, T. 1995. Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception,* **24**: 1075–81.

Crick, F. 1995. *The Astonishing Hypothesis: The Scientific Search For The Soul*. New York: Scribner.

Damasio, A. and Damasio, H. 2006. Minding the body. *Daedalus* 135, No. 3, 15-22.
Dennett, D. 1991. *Consciousness Explained*. Boston: Little, Brown and Co.
Dolan, R. 2006. The body in the brain. *Daedalus* 135, No. 3, Pages 78-85
Edelman, G. 2006. The embodiment of mind. *Daedalus* 135, No. 3, 23-32
Gallagher, S. 2005. Metzinger's matrix: Living the virtual life with a real body. *Psyche* 11 (5). *www.theassc.org/files/assc/2614.pdf*
Gallagher, S. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences* 4 (1): 14-21.
Gazzaniga, M. 1998. *The Mind's Past.* Berkeley: University of California Press.
Gillihan SJ, and Farah MJ. 2005. Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychol Bull.* 131(1), 76-97.
Giummarra MJ, Gibson SJ, Georgiou-Karistianis N, Bradshaw JL. 2008) Mechanisms underlying embodiment, disembodiment and loss of embodiment. *Neurosci Biobehav Rev* 32: 143–160
Graziano M. and Botvinik M. 2001. How the brain represents the body: insights from neurophysiology and psychology. In: W. Prinz and B. Hommel (eds.), *Common mechanisms in perception and action, attention and performance* XIX. (136–157). Oxford: Oxford University Press.
Jackson, S. R. Buxbaum, L. and Coslett, H. B. (Eds). Forthcoming. The body in the brain: Body representations, processes and neural mechanisms. Special Issue of *Cognitive Neuroscience*.
James, W. 1890. *The Principles of Psychology*. New York: Dover, 1950.
Legrand D. & Ruby P. 2009. What is self specific? A theoretical investigation and a critical review of neuroimaging results. *Psychological Review*, 116 (1) 252–282.
Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8: 529-66.
Libet, B., Gleason, C. A., Wright, E. W. and Perl, D. K. 1983. Time of conscious intention to act in relation to cerebral activities (readiness potential): The unconscious initiation of a freely voluntary act. *Brain* 106: 623-42.
Melzack, R. 1990. Phantom limbs and the concept of a neuromatrix. *Trends in Neuroscience* 13: 88-92.
Metzinger, T. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.
Metzinger, T. 2005. Precis: Being No-One. *Psyche* 11 (5). http://www.philosophie.uni-mainz.de/metzinger/publikationen/precis.pdf
Neisser, U. 1988. Five kinds of self-knowledge. *Philosophical Psychology*, 1: 35-59.
Noë, A. 2002. Is the visual world a grand illusion? *Journal of Consciousness Studies*, **9** (5–6): 1–12
Pöppel, E. 1978. Time perception. In R. Held, H.W. Leibovitz and H. L. Teuber (eds.), *Handbook of Sensory Physiology,* Vol. 8. New York: Springer.
Putnam, H. 1992. Brains in a Vat. In K. DeRose and T.A. Warfield (eds.), *Skepticism: a Contemporary Reader*. Oxford: Oxford University Press.
Schechtman, M. 2011. The narrative self. In S. Gallagher (ed.), *The Oxford Handbook of the Self* (394-416). Oxford: Oxford University Press.
Strawson, G. 1999a. The self. In S. Gallagher and J. Shear (eds.), *Models of the Self* (1-24). Exeter: Imprint Academic.

Strawson, G. 1999b. The self and the SESMET. In S. Gallagher and J. Shear (eds.), *Models of the Self* (483-518). Exeter: Imprint Academic.
Vogeley, K. and Gallagher, S. 2011. Self in the brain. In S. Gallagher (ed.), *Oxford Handbook of the Self* (111-136). Oxford: Oxford University Press.
Wegner, D. 2003. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.