The Brain and Its Self[1]

PATRICIA SMITH CHURCHLAND
Philosophy Department
University of California, San Diego

SNAKING OUT of the MRI tube where I had lain for the past half hour, I glimpsed Dr. Hanna Damasio studying the lab's display screen. Off the gurney, I went to her side and stared at the image of my brain on the screen. "Is that—*me?*" Well, yes, in a certain sense. And yet not simply, or merely, me. Certainly not *familiarly* me. Here is what I thought: "Somehow, starting in infancy, my brain built a story about itself—its body, its history, its 'now,' and its world. From the inside, I know that story—though I think of it as reality, not just a story. Indeed, it *is* my inner reality. So how does this happen? What is it for me to be a construction of my brain?"

In one form or another, these questions have had a long and convoluted history, born in the unflinching curiosity of the ancient Greeks, and finding voice in diverse cultures. Until recently, the only explanatory resources for addressing puzzling behavior depended on mythologizing in the case of others, and myth-filtered introspection in the case of oneself. Not surprisingly, early explanations invoked possession by devils or, if you were luckier, divine forces, to account for epileptic seizures or schizophrenic hallucinations. In the absence of understanding, the punishment theory of mental dysfunction commanded widespread belief, yet it was wholly untestable—and essentially untested.

Melancholia (what we now call chronic depression) and phobias were often surmised to be essentially character flaws—flaws that might be overcome with sufficient gumption. The existence of witches, hexes, curses, and spells had a far longer history as brute fact than does our appreciation of such potent neurochemicals as serotonin. Obsessive handwashing, a mere fifteen years ago, was widely assumed to be a manifestation of repressed sexuality. Nevertheless, even as early as 400 BC, the great Greek physician, Hippocrates, was convinced that events such as

[1] Read 24 April 2009, as part of the symposium "The Relation of Organ, Limb, and Face Transplantation." This article is based on chapter 3, "Self & Self-Knowledge," in *Brain-Wise* (Cambridge, Mass.: MIT Press, 2002).

sudden paralysis or creeping dementia had their originating causes in brain damage—which implied, in his view, that normal movement and normal speech had *their* originating causes in the well-tempered brain.

Brains, however, are not easy organs to figure out. Imagine Hippocrates, dissecting the brain of a dead warrior after autopsy, and pondering an area of sword-destroyed tissue. To what theoretical resources could he reach to begin to make sense of something so complex as the relation between fluent speech and the pinkish tissue found in the skull? Remember, in 400 BC nothing was understood about the nature of the cells that make up the body, let alone of the special nature of the cells that make up the brain. Techniques for isolating neurons—brain cells—to see what they looked like could begin only in the nineteenth century. Techniques for isolating *living* neurons to explore their *function* did not appear until well into the twentieth century.

Figuring out how neurons do what they do requires high-level technology. And that, needless to say, depends on immense infrastructural science; on cell biology, advanced physics, and twentieth-century chemistry. It requires sophisticated modern notions like molecule and protein, and modern tools like the light microscope and the electron microscope.

What is most important, making progress on how brains work depended on understanding electricity. This is because what makes brain cells special is their capacity to signal one another by causing fast micro changes in each other's electrical state. Living as we do in an electrical world, it is sobering to recall that as late as 1800, electricity was typically considered deeply mysterious and quite possibly occult. Only after discoveries by Ampère and Faraday at the dawn of the nineteenth century was electricity clearly understood to be a physical phenomenon, behaving according to well-defined laws, and capable of being harnessed for practical purposes.

In this century, modern neuroscience and psychology allow us to go beyond myth and introspection to approach the "self" as a natural phenomenon whose causes and effects can be addressed by science. Helped by new experimental techniques and new explanatory tools, we can pry loose real understanding of how the brain comes to know its own body, how it builds a coherent model of its world, and how changes in brain tissue can entail changes in the very self itself. Neurobiology is beginning to reveal why some brains are more susceptible than others to alcohol or heroin addiction, or why some brains slide into incoherent world-models. Progress is visible on the staged emergence of self in childhood, as well as the cruel inch-by-inch loss of self in dementia. Though well short of full answers, neuroscience has discovered much about the effects of localized brain lesions on complex decision making or speech or voluntary behavior.

All these developments are part of the story of the neuroself. True enough, neuroscience has not advanced enough to yield complete and detailed answers to the whole range of questions. Perhaps some questions will forever exceed the neurobiological reach, though it may be hard to tell whether such problems are just "as yet unsolved" or whether they are truly unsolv*able*. In any case, incomplete but powerful answers anchored in data can often provide a foothold for the next step. And then, in turn, for the next step thereafter. That is how science proceeds —one step at a time.

As I watched the computer monitor showing my brain tilted at various angles and cut at various slices, what stirred was the idea that I might come to know my neuroself at least as well as I know my psycheself. Or, at least, someone in the next generation might. I imagined Hippocrates, looking at the image of *his* brain, agog with excitement and eager to experiment further.

1. Isn't "The Brain-Made Self" Paradoxical?

The question "How does my brain make myself?" does have a sort of "snake-with-tail-in-mouth" quality. To evade the paradox, I adopt a pair of pragmatic principles: (1) ask instead, "How does *a* brain make *a* self?" thereby putting the paradox at arm's length, and (2) follow the facts first, and let the paradoxes fend for themselves. As a neurophilosopher, I predict that the paradoxes might well vaporize once the neuroscience gets a bit clearer.

This is not just wishful thinking. I had seen other ostensible paradoxes about the world dissolve as so much candy floss in a flame. Once the relevant science revealed the reality behind the mystifying appearances, what had seemed counterintuitive came to be familiar and largely obvious. The idea that the Earth moves or that living things are made of dead molecules—each lost its patina of paradox in the gentle light of experiment and explanation. My hunch is that this could happen here, too. Consequently, confronting the wheelie snake is something I gladly defer until the science is a bit further along.

2. What Kind of Thing Is the Self?

It is this question that makes philosophers of us all, at some time or other. Not everyone wonders late into the night what the stars are made of or how the immune system works. But wondering what it is that makes me *me*, is never very far from one's elbow. Philosophers since Plato in the fifth century BC have tried to make progress in coming up with satisfying answers—or, more minimally, with ways of structuring the

question to avoid spiraling down into confusion. The great eighteenth-century Scottish philosopher David Hume gave the questions their clearest analysis and set the stage for modern scientific investigation.

Hume came to the quite shocking realization that if you look inwardly to try to observe a distinctly "self" thing, there does not seem to be any self thing there to perceive. What there is, is a continuously changing *flux* of visual perceptions, sounds, smells, emotions, thoughts, and so forth. Amongst all those, however, there does not exist a single, continuous "felt" sensation that one can attend to and say, "That's the self," as one can attend to a felt sensation and say, "That's a pain."

Yet nothing could be more evident than that there seems to be a single thread of "me-ness" that runs through the entire fabric of experience. We all have a robust and undeniable self-representation. We generally awake from a deep sleep knowing who we are, even if we are confused about when and where we are. Normally, we do not doubt that "I am essentially the same person today as I was yesterday, and the day before that." Normally, we know without pausing to figure it out that "this body is my own"—that "this hand and this foot are both parts of my body." When I talk to myself about becoming a marathon runner, I know that it is *me* talking to *myself*. We know very well that if we fail to plan for future contingencies, our future selves may suffer, and we care *now* about that *future* self.

Here is Hume's conundrum: I know myself about as well as I know anything, yet my self is not anything that I can ever observe—at least not in the way that I can observe touches or warmth or fatigue. The dilemma can be put this way: on what is the idea of the self based, if not on a continuous sensation? If it is an abstract kind of thing—not an *observable* kind of thing—what are its properties and where does it come from?

As neuroscience and experimental psychology have progressed in this century, an updated version of Hume's problem has emerged: how does the brain—a network of trillions of cells—generate this representation of a unitary *self*? What are the neural mechanisms underlying such self-representation? One important source of information will be pathologies involving self-representation. Some insight can be garnered from stroke patients who deny that these hands are in fact *their* hands, or who have lost all memory of life events before the brain damage or who feel that they have lost their "will." Schizophrenics or patients on the anesthetic ketamine often suffer "depersonalization" feelings—feelings that they are dead or possessed. These phenomena, too, indicate how the "self-evidence" of the self is underpinned by complex brain activity. By revealing the fracture lines of the self-representation, these kinds of cases allow us to see what is well hidden in the normal case.

3. Brains Emulate Body and Self

A. The General Idea

Referring to "the self" suggests the self must be a kind of *thing*, such as a specific organ in the brain, the way that the spleen or the pancreas is a specific organ in the body. Clearly, however, the "pancreas paradigm" for thinking about the self won't work. The self is not an organ in the brain; nor, so far as we know, is there a discrete region of the brain that "makes" the self. But if the self is not a thing like the pancreas, and if it is not a continuous sensation, what is it?

The best hypothesis is that it involves a complex idea (representation) that the brain generates through activity in various different regions, including the regions representing the body and a representation using memory of the past. The brain activity that we know introspectively as "myself" is probably part of a set of larger patterns of activity the brain deploys for making sense of and getting by in its world. Given these considerations, it is preferable to talk about the problem of *self-representation* rather than the problem of the *self*. But what is it for the brain to represent *anything*, let alone "self"? Must there not *be* a self if the brain represents it?

B. Representation in the Brain

Part of the major business of nervous systems, from crayfish to humans, is to make good predictions about food, mates, enemies, and friends, so that the body can live on to reproduce. Poor predictors often end up as meals for better predictors. Imposing structure on our sensory stimuli in the service of better prediction is what representation is all about. Using internal representations allows for much more sophisticated behavior than mere stimulus-response reflexes. Using internal representations is a common strategy that nervous systems have developed as part of evolution's way of favoring adaptive structures.

The philosopher Rick Grush[2] has developed a useful tool for getting a handle on this. Suppose I am running a huge construction crane, which is a very high-tech crane that I can operate from the comfort of my office a mile away. It would be a good idea for the engineer to design it so that I have access to a small-scale model that shows where the hook will be if I give the order for a certain movement. That way I can correct my movement without waiting for feedback from the gigantic

[2] See Rick Grush, "Emulation and Cognition" (Ph.D. diss., UCSD, 1995); idem, "The Architecture of Representation," *Philosophical Psychology* 10 (1997): 5–25.

hook-in-the-world. The emulator in my office generates internal feedback that helps me predict. Even better, the designer could allow me to fiddle with the model so that I can test possible movements before I choose the best, thereby maximizing the accuracy of the movement when I do finally make the actual hook move. Very crudely, this is what Grush thinks brains do. They build "emulators" of the world and of their bodies in that world.

Of course if you looked in my brain you would not see a miniature world of tiny trees and dogs and so forth—just cells connected to cells, signaling each other and displaying patterns of activity. Nor is there a little person in my head who sits and watches a screen. That part of the emulator story does not at all fit what brains do. What we can take from the emulator story is the similarity in function. Some patterns of neuronal activity seem to be performing the same function as the crane-emulator.

Exactly how this works is not known. Nevertheless, it seems evident that inner modeling of the body and its world is an evolutionary achievement that means the organism can do smarter things than otherwise. Not all aspects of the organism's world need be emulated in its brain—only those that matter to it, given its way of making a living.[3] Bees can detect ultraviolet light, and that helps them forage among flowers. Humans do not perceptually represent that aspect of the world, unless we build a tool to do it for us. In a similar vein, I shall not need all aspects of the crane-world explicitly emulated—just those relevant to getting the job done.

FIGURE 1. A Dalmatian dog in the dappled light, standing on a leaf-strewn pavement

Some world-emulation will be on-line, as when the brain displays perceptual construction and filling-in. Thus we see a Dalmatian in the leafy background even though the stimulus itself is degraded (fig. 1). We see the tomato as uniformly red even though it is shadowed and highlighted and partially occluded; we hear our names spoken in a noisy room. Off-line, so to speak, we remember where we cached the food by the river; we plan how to cross a turbulent stream; we daydream and fantasize.

[3] Kathleen Akins, "Of Sensory Systems and the 'Aboutness' of Mental States," *Journal of Philosophy* 93.7 (1996): 337–72.

On the Grush hypothesis, the brain emulates the ecologically relevant —the "relevant-to-my-kind-of-creature"—features of the world, and then manipulates these emulations to plan, hide, forage, and so forth. I may consider the problem of crossing the turbulent stream, go on to imagine a route that would be easier if a log were stretched from one rock to another, and go on to imagine the size of the log needed and how to get it into place. This involves manipulation of the image or, as we may say, of the river-crossing emulation.

C. Body Models

So far we have focused on emulations that capture features of the outside world, but brains can also emulate aspects of the body. You can, for example, imagine your body standing when you are sitting, or the size it was when you were five. Sexual fantasies are potent instances in which real body effects can be produced by the brain's manipulation of a two-body emulation. Imaginary tennis and golf have been demonstrated to be highly significant in improving one's actual game.

Hiding your body from another viewer requires enormous representational sophistication. You need some understanding of the visual aspect of your body, its proportions relative to the shield. Most critically, you need to grasp how the scene will look from *another's* viewpoint. Remember playing hide-and-seek, and the importance of knowing the visibility of one's body from various perspectives. From the perspective of whoever is "it," there must be no feet sticking out, no hair showing above, though visibility to our fellow hiders doesn't matter.

A very young child may think she is hidden from others when she puts her hands over her eyes. She does not yet have a representation of how she looks from another's eyes. But she probably has spent lots of time watching her fingers manipulate food, toys, the dog, and her own toes, and probably her visually-anchored body-schema is still emerging. An integrated body-schema, with both visual and somatosensory dimensions, will have begun to develop from her very early days, even if she cannot yet manage all the subtleties of the difference between "I can see me" and "You can see me."

D. Self-Models

Additionally, complex brains can emulate aspects of what the brain itself is doing, and the "self," I suggest, is one such result. That is, it may have a model of the brain's activities, perhaps cast in perceptual images resembling familiar external events. As the philosopher Patricia Kitcher sees it, something like this is what Kant had in mind as the basis for "self."

The metaphors in common use give some inkling of how those emulations are structured. When faced with a difficult choice, people refer to inner struggles or tug-of-war games; forgetting may be likened to the fading of print or blocking by a barrier. Desires may be said to overpower one, or to have a grip or a hold; they may possess one and one may surrender to them. They may be repressed ("pushed below the surface") only to bob up in a new disguise. Fears can run away with or dominate oneself; knowledge is seeing; hope can spring eternal in the human breast. And so on.

In short, the hypothesis I favor is that the self is a kind of emulation, constructed by the brain, for integrating and making sense of the inner world of the brain in its relation to the external world, including the other-person-world. Minimally, it has (1) a body component, (2) a "what-I-am-aware-of-now" component, (3) a stable but modifiable background of preferences, habits, skills, temperament, and so forth, and (4) a memory-based autobiographical component.[4] These components are interrelated, but are also, to some extent, dissociable.

This is obviously not a precise characterization. For that, we need to understand much more about the details of brain function at many levels of organization, from the single cell to the whole brain. The notion of "representation," like the notion of an emulator, is more like a place-holder waiting for a detailed theory of brain function than a precise term in a well-developed theoretical framework. Nevertheless, we do have some clear ideas on the general role it needs to fill, and the important task will be to find experiments to test the hypothesis.[5]

4. "How can you have any self-esteem if you think you are just a piece of meat?"

So asked a forthright student. My answer is that first, brains are not *just* pieces of meat. The human brain is what makes humans capable of painting the Sistine Chapel, of designing airplanes and transistors, of skating and reading and playing Chopin. To that degree, it is a truly astonishing and magnificent kind of "wonder-tissue," as Dan Dennett puts it. Whatever self-esteem justly derives from our accomplishments does so *because* of the brain, not in spite of it.

Second, if we thought of ourselves as glorious creatures before we knew that the brain is responsible, why not continue so to feel after the

---

[4] See also Owen Flanagan, *Consciousness Reconsidered* (Cambridge, Mass.: MIT Press, 1992); Arnold Ludwig, *How Do We Know Who We Are?* (Oxford, New York: Oxford University Press, 1997); Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991).

[5] Patricia Kitcher, *Kant's Transcendental Psychology* (Oxford, New York: Oxford University Press, 1990).

discovery? Why does the knowledge not make us more interesting and remarkable, rather than less so? We can be thrilled by the spectacle of a volcano erupting or a calf being born or bone healing before we understand what volcanoes are and how reproduction or healing works. Being the creatures we are, however, commonly we are even more thrilled in the embrace of the knowledge about volcanoes and birth and bones.

5. Concluding Remarks

Self-representation in humans is a highly complex business. Richly layered in language, embedded in a social context, and backed by a detailed, if selective, autobiographical history, it is also replete with the qualitative features of conscious experience. We talk to ourselves: "Do I really want a cigarette?"; "Why did I allow myself to get angry?" We think about ourselves in culturally shared metaphors: "I hid my true self from myself "; "I lost control of myself"; "I struggled to keep myself from falling apart." We have important self-regulating feelings not directly linked to a particular sensory modality: we can feel comfortable, uneasy, unfamiliar, confident, ashamed, embarrassed, and so on.[6] The neurobiology of consciousness is a central topic that ideally should be discussed in this context. Limitations of space, however, make that a topic for another occasion.

Much—so *very* much—remains to be discovered, and my theoretical resources are limited to drawing on where cognitive neuroscience is now. There are few puzzles about the brain that we can say are flat-out solved, and I have too much of the farm in me to count hens in advance of the hatch. Even so, it is well known that discoveries in the last three decades have allowed us new insights undreamt of in our philosophy, and it will be surprising if more are not to come. Will we come to think of ourselves in a different light?

The retrofitting of time-honored ideas is already visible, and some questions now routinely researched by graduate students were unconceived of a mere twenty years ago. What further changes can be expected, or in what directions, is anybody's guess. From where I stand now, it seems to me likely that our understanding of what it is to be "in control" of one's behavior, of what consciousness, personality, and character are, will change, perhaps quite profoundly. As with many other developments in human intellectual history, I expect there will be struggles between superstition and science, between the old and familiar on

---

[6] Antonio Damasio, *Descartes' Error* (New York: Putnam and Sons, 1994); P. S. Churchland, "Feeling Reasons," in *Neurobiology of Decision-Making*, ed. Antonio Damasio, Hanna Damasio, and Yves Christen (Berlin, New York: Springer-Verlag, 1996).

the one hand, and the new and unfamiliar on the other. Just as cell biology and molecular biology achieved humanizing results by overturning demonic-possession and punishment theories of disease, so I predict neuroscience will have humanizing effects as it reveals more of what it is that makes us human.

Part of what makes science so intriguing is the unpredictability of how things will look after the next bend in the river. Intriguing, too, is the creation of new thought-tools provoked by totally unpredictable results but needed to get to the heart of the puzzle. Scientifically, we are the lucky ones. We are alive as the twentieth century—the greatest century for science—gives birth to the twenty-first. We have a chance to follow the river, and find out what really is the lay of our land.