My body has a mind of its own

Daniel C. Dennett

Tufts University

Center for Cognitive Studies

Medford MA 02155

In life, what was it I really wanted? My own conscious and seemingly indivisible self was turning out far from what I had imagined and I need not be so ashamed of my self-pity! I was an ambassador ordered abroad by some fragile coalition, a bearer of conflicting orders, from the uneasy masters of a divided empire. . . . As I write these words, even so as to be able to write them, I am pretending to a unity that, deep inside myself, I now know does not exist."

--William Hamilton, 1996, p134

"Language was was given to men so that they could conceal their thoughts."

-- Charles-Maurice de Talleyrand

"My body has a mind of its own!" Everybody knows what this exclamation means. It notes with some surprise the fact that our bodies can manage many of their key projects without any conscious attention on our part. Our bodies can stride along quite irregular ground without falling, avoiding obstacles and grabbing strategic handholds whenever available, pick berries and get them into the mouth or the bucket with little if any attention paid, and–notoriously–initiate preparations for sexual activity on a moment's notice without any elaborate decision-making or evaluation discernible, to say nothing of the tight ship run by our temperature maintenance system and our immune system. My body can keep life and limb together, as we say, and even arrange for its own self-replication without any attention from me. So what does it need *me* for?

This is another way, perhaps a better way, of asking why consciousness (our human kind of consciousness, at least) should evolve at all. The sort of consciousness (if it *is* a sort of consciousness) that is manifest in alert and timely discriminations for apt guidance of bodily trajectory, posture, and resource allocation is exhibited, uncontroversially, by invertebrates all the way down to single-celled organisms, and even by plants.1 Since the self-protective dispositions of the lobster can be duplicated, so far as we can tell, in rather simple robots whose inner states inspire no conviction that they must "generate phenomenology" or anything like it, the supposition that nevertheless there must be "something it is like to be" a lobster begins to look suspicious, a romantic overshooting of anthropomorphism, however generous-spirited. If a lobster can get through life without a self (or very much of a self), why should it have a 'selfy' self (Dennett, 1991)? Maybe it doesn't. Maybe it (and insects and worms and . . . .) are "just

[1]This robotic sensitivity-*cum*-action-guidance is what I used to call *awareness*$_2$, to distinguish it from the reportable kind of consciousness that might be only a human gift, *awareness*$_1$ (Dennett, 1969), a pair of awkward terms that never caught on, though the idea of the distinction is still useful, in my opinion.

robots". The problem with this supposition is that we really don't know that *no* robot could be conscious, so we shouldn't be confident that what is apparently true of *simple* robots would be true of all possible complex robots. After all, we are conscious, and if materialism is true, we are made of nothing but robots–cells–and is such a complex not itself a robot?2 If we want to have a way of expressing our supposition that lobsters, say, are "mere" automata, we need to anchor that "mere" to some upper bound on complexity. Here is one way. In 1991 (p310ff), in response to some imagination-blockades that affect philosophers thinking about the entirely imaginary phenomenon of zombies, I proposed distinguishing a subspecies of zombies, *zimboes*, which unlike their simpler brethren are *reflective*, capable of higher-order self-monitoring. Whether or not some very complex zimboes are conscious, non-zimbo zombies–perhaps we might call them *zombots* to highlight their simplicity–meet our intuitive demands for being obviously not conscious (if thermostats and cell phones are obviously not conscious), so now we can rephrase our supposition as the hypothesis that lobsters are apparently just zombots. And then we can ask ourselves which other genera are also zombots, and which are zimboes, whether or not they are conscious like us. If consciousness is not just zimbohood, we might nevertheless use the zimbo-zombot distinction as our temporary scaffolding in our search for consciousness. Suppose this enabled us to distinguish two quite different styles of nervous system: the simple, relatively low-priced zombot arrangement, lacking central self-monitoring capabilities, and the more expensive and sophisticated zimbo arrangement, capable of significant varieties of higher-level self-

[2]I find it strategically useful to insist that individual cells, whether prokaryotic, archaic, or eukaryotic, are basically robots that can duplicate themselves, since this is the take-home message of the last half century of cell biology. No more romantic vision of living cells as somehow transcending the bounds of nanobothood has any purchase in the details of biology, so far as I can see. Even if it turns out that quantum effects arising in the microtubules that criss-cross the interiors of these cells play a role in sustaining life, this will simply show that the robotic motor proteins that scurry back and forth on those microtubules are robots with access to randomizers.

monitoring. As we learned more about the adroitness and versatility and internal organization of spiders or octopuses (or bats or dolphins or bonobos), we might uncover some further impressive thresholds that persuaded us to grant consciousness (of the impressive kind–whatever that means) to these creatures, but for the time being, we are the only species that everybody confidently characterizes as conscious is ours. Our confidence is grounded in the simple fact that we are the only species that can compare notes.

Consider a remark by Daniel Wegner:

> We can't possibly know (let alone keep track of) the tremendous number of mechanical influences on our behavior because *we* [italics added--DCD] inhabit an extraordinarily complicated machine (2002, p27)

Wegner presumably wouldn't have written this if he hadn't been comfortable assuming that we all know what he is talking about, but just who is this *we* that 'inhabits' the brain? There is the Cartesian answer: each of us has an immortal, immaterial soul, the *res cogitans* or thinking thing, the seat of our individual consciousness. But once we set that answer aside–as just about everybody these days is eager to do–just what thing or organ or system could Wegner be referring to? My answer, compressed into a slogan by Giulio Giorelli, who used it as a headline for an interview with me in *Corriere della Serra* in 1997, is this: *Si, abbiamo un anima. Ma è fatta di tanti piccoli robot.* Yes, we have a soul. But it's made of lots of tiny robots. Somehow, the trillions of robotic (and unconscious) cells that compose our bodies organize themselves into interacting systems that sustain the activities traditionally allocated to the soul,

the Ego or self. But since we have already granted that *simple* robots are unconscious (if toasters and thermostats and telephones are unconscious), why couldn't teams of such robots do their fancier projects without having to compose *me?* If the immune system has a mind of its own, and the hand-eye coordination circuit that picks the berries has a mind of its own, why bother making a super-mind to supervise all this?

George Ainslie notes the difficulty and has a suggestion:

> Philosophers and psychologists are used to speaking about an organ of unification called the 'self' that can variously 'be' autonomous, divided, individuated, fragile, well-bounded, and so on, but this organ doesn't have to exist as such. (2001, p43)

How could this be? How could an organ that doesn't have to exist *as such* exist at all? And, again, *why* would it exist? Another crafty thinker who has noticed the problem is the novelist Michael Frayn, whose narrator in *Headlong* muses:

> Odd, though, all these dealings of mine with myself. First I've agreed a principle with myself, now I'm making out a case to myself, and debating my own feelings and intentions with myself. Who is this *self,* this phantom internal partner, with whom I'm entering into all these arrangements? (I ask myself.) (Frayn, 1999, p143)

Although Frayn may not have intended to answer his own question, I think he has in fact

provided the key to the answer in his parenthesis: “I ask myself.” It is only when *asking* and *answering* are among the projects undertaken by the teams of robots that they have to compose a virtual organ of sorts, an organ that “doesn’t have to exist as such” --but does have to exist.

This, in any case, has been suggestively argued by the ethologist and roboticist David McFarland (1989) in a provocative, if obscure, essay. According to McFarland, “Communication is the only behavior that requires an organism to self-monitor its own control system.” I’ve been musing over this essay and its implications for years, and I still haven’t reached a stable conviction about it, but in the context of the juxtaposition of Wegner and Ainslie and the other excellent speakers in Birmingham, I think it is worth another outing.

Organisms are correctly seen as multi-cellular communities sharing, for the most part, a common fate (they’re in the same boat). So evolution can be expected to favor cooperative arrangements in general .Your eyes may, on occasion, deceive you–but not on purpose! (Sterelny, 2003) Running is sure to be a coordinated activity of the limbs, not a battle for supremacy. Nevertheless, there are bound to be occasions when subsystems work at cross purposes, even in the best-ordered communities of cells, and these will in general be resolved in the slow, old-fashioned way: by the extinction of those lineages in which these conflicts arise most frequently. The result is control systems that get along quite well *without* any internal self-monitoring. The ant colony has no boss, and no virtual boss either, and gets along swimmingly with distributed control that so far as we can tell does not engage or need to engage in high level self-monitoring. According to McFarland, organisms can very effectively control themselves by a collection of competing but ‘myopic’ task-controllers that can interrupt each other when their

conditions (‘hunger’ or need, sensed opportunity, built-in priority ranking, . . .) outweigh the conditions of the currently active task controller. Goals are represented only tacitly, in the feedback loops that guide each task-controller, but without any global or higher-level representation. (One might think of such a task-controller as “uncommented” code–it works, but there is nothing anywhere in it that can be read off about what it does or why or how it does it.). Evolution will tend to optimize the interrupt dynamics of these modules, and nobody’s the wiser. That is, *there doesn’t have to be anybody home* to be wiser!

But communication, McFarland thinks, is a behavioral innovation that changes that. Communication requires a central clearing house of sorts *in order to buffer the organism from revealing too much about its current state* to competitive organisms. In order to understand the evolution of communication, as Dawkins and Krebs (1978) showed in a classic article, we need to see it as *manipulation* rather than as purely cooperative behavior. The organism that has no poker face, that communicates state directly to all hearers, is a sitting duck, and will soon be extinct. What must evolve instead is a communication-control buffer that creates (1) opportunities for guided deception, and coincidentally (2) opportunities for self-deception (Trivers, 1985), by creating, for the first time in the evolution of nervous systems, explicit (and more “globally” accessible) *representations* of its current state, representations that are detachable from the tasks they represent, so that deceptive behaviors can be formulated and controlled. This in turn opens up structure that can be utilized in taking the step, described in detail by Gary Drescher (1991), from simple *situation-action machines* to *choice machines*, the step I describe as the evolutionary transition from Skinnerian to Popperian creatures.(Dennett, 1995).

I wish I could spell all this out with the rigor and detail that it deserves, but I have been unable to make much progress on this important task in the the time available. The best I can do at this point is simply gesture in the directions that strike me as theoretically promising and encourage others to mine this fine vein. What follows are some informal reflections that might contribute.

Consider the chess-playing computer programs that I so often discuss. They are not conscious, even when they are playing world-class chess. There is no role for a *user-illusion* within them because, like McFarland's well-evolved non-communicators, they are more or less optimized to budget their time appropriately for their various subtasks. Would anything change if the program were enabled/required to communicate with others–either its opponent or other kibitzers? Some programs now available have a feature that permits you to see just which move they are currently considering (see, e.g., http://chess.captain.at/) but this is not communication; this is involuntary self-exposure. a shameless display that provides a huge source of valuable information to anyone who wants to try to exploit it. In contrast, a program that could consider its communications as informal moves–social ploys-- in the enlarged game of chess–the game that some philosophers (e.g., Haugeland, 1998) insist is real chess, unlike the socially truncated game that programs now play--would have to be able to "look at" its internal states the same way a poker player needs to look at his cards in order to decide what action to take. "What am I now trying to do, and what would be the effect of communicating information about that project to this other agent?" (It asks itself) In other words, McFarland imports Talleyrand's cynical dictum about language and adapts it to reveal a deep biological truth: explicit self-monitoring was invented to conceal our true intentions from each other while permitting us to reveal strategic versions of

those intentions to others.

If this is roughly right, then we can also see how this capacity has two roles to play: export and import. It is not just that we can use communication to give strategic information about what we are up to, but we can put up to things by communication. “A voluntary action is something a person can do when asked.” (Wegner, p32) The capacity to respond to such requests, whether initiated by others or by oneself, is a capacity that must require quite a revolutionary reorganization of cerebral resources.3 This prospect is often overlooked by researchers eager to stress the parallels and similarities between human subjects and animal subjects when they train a monkey (typically) to ‘indicate’ one thing or another by moving their eyes to one or another target on screen, or to press one of several buttons to get a reward. A human subject can be briefed in a few minutes about such a task, and thereupon, with only a few practice trials, execute the instructions flawlessly for the duration of the experiment. The fact that preparing the animal to perform the behavior usually involves hundreds or even thousands of training trials does little to dampen the conviction that the resulting behavior counts as a “report” by the animal of its subjective state.4 But precisely what is missing in these experiments is any ground for believing that the animal *knows it is communicating* when it does what it does. And if it is *not* in the position of an agent that has decided to tell the truth about what is going on in it now, there is really no reason to treat its behavior as an intentional informing. Its carefully sculpted actions

[3]Thomas Metzinger, *Being No One*, 2003, has some excellent suggestions about what he calls the *phenomenal self-model* and its revolutionary capacities.

[4]See *Sweet Dreams*, p169-70 for earlier remarks on this. I myself have given more credence in the past to this proposal than I now think appropriate. See “What is it like to be a bat?” in CE, pp8, esp446ff.

may *betray* its internal state (the way the chess program willy-nilly divulges its internal state), but this is not the fruit of self-monitoring.

If something along these lines is right, then we have some reason to conclude that contrary to tradition and even "common sense", there is scant reason to suppose that it *is* like anything to be a bat. The bat's body has a mind of its own, and doesn't need a "me" to inhabit it at all. Only we who compare notes (strategically) inhabit the complicated machines known as nervous systems.

References

Ainslie, George, 2001, *Breakdown of Will*, Cambridge: Cambridge Univ. Press.

Dennett, Daniel C., 1969, *Content and Consciousness,* London: Routledge & Kegan Paul

----1991, *Consciousness Explained*, New York: Little Brown

----2005, *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, Cambridge, MA: MIT Press.

Drescher, Gary, 1991, *Made-Up Minds: A Constructivist Approach to Artificial Intelligence,* MIT Press.

Frayn, Michael, 1999, *Headlong*, London: Faber & Faber.

Hamilton, William, 1996, *Narrow Roads of Gene Land,* Vol 1. Oxford, W. H. Freeman

Haugeland, John, 1998, *Having Thought: Essays in the Metaphysics of Mind,* Cambridge, MA: Harvard Univ. Press

McFarland, David, 1989, "Goals, No-Goals and Own Goals," in Alan Montefiore and Denis Noble, eds., *Goals, No-Goals and Own Goals: A Debate on Goal-Directed and Intentional*

Metzinger, Thomas, 2003, *Being No One*, 2003, Cambridge, MA: MIT Press.

Sterelny, Kim, 2003, *Thought in a Hostile World: The Evolution of Human Cognition*, Oxford: Blackwell.

Trivers, Robert, 1985, *Social Evolution*, Benjamin/Cummings.

Wegner, Daniel, 2002, *The Illusion of Conscious Will*, Cambridge, MA: MIT Press.