The Self as a Responding—and Responsible—Artifact

DANIEL C. DENNETT

Center for Cognitive Studies, Tufts University, Medford, Massachusetts 02155, USA

ABSTRACT: The powerful illusion of a unified, Cartesian self responsible for intentional action is contrasted with the biologically sounder model of competitive processes that yield an only partially coherent agency, and the existence of the illusion of self is explained as an evolved feature of communicating agents, capable of responding to requests and queries about their own decisions and actions.

KEYWORDS: self; consciousness; agency; free will; timing of voluntary actions

It *seems* to us as if each of us *is* a self, a unified, rational agent, in control of a body. I am in charge of my body and you are in charge of your body. Descartes, in the 17th century, famously called this agent-in-charge a *res cogitans*, a thinking thing, and deduced to his own satisfaction that it could not be a physical, mechanical thing, but rather an immaterial thing, which he could conveniently identify with his immortal soul. (It was convenient, given his Roman Catholicism.) Today, materialism has swept dualism and its insoluble mysteries of interaction aside, so this is no longer regarded as a convenient, or even tenable, hypothesis.

Efforts to identify the self—a mortal and material soul, you might say—with a particular subsystem in the brain run into snags at every turn. I call this the fallacy of the Cartesian Theater, the place in the brain where it all comes together for conscious appreciation and decision. This just cannot be right. All the work done by the imagined homunculus in the Cartesian Theater must be distributed around to various lesser agencies in the brain. But some people just hate this idea. They think, mistakenly, that unless there is a Cartesian

Address for correspondence: Daniel C. Dennett, Center for Cognitive Studies, Tufts University, Medford, MA 02155-3952. Voice: 617-627-3297; fax: 617-627-3952.
ddennett@tufts.edu

Ann. N.Y. Acad. Sci. 1001: 39–50 (2003).
doi: 10.1196/annals.1279.003

Theater, there is no consciousness. They are wrong for many reasons, not least of which is that the *specs* for such a organ or faculty of the body turn out to be, shall we say, optimistic. We are just not that unified. The late, lamented evolutionary biologist William Hamilton, reflecting on his own uneasiness with his recognition of this fact, put the issue particularly well:

> 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. (Hamilton, 1996, p. 134)

Such coherence as we manage in our lives is the dynamic and shifting resultant of competitions between all manner of structures, some of them deeply rooted in our genes (and with our paternally and maternally inherited genes in something of a permanent tug-of-war) and some of them temporary *ad hoc* coalitions of "interests" engaged in "intertemporal bargaining" (in the terms of George Ainslie, 2001), a recursive and self-referential process on a hair-trigger.

> 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. (Ainslie, 2001, p. 43)

Given the literal chaos brewing in our brains, given the manifest absence of any King Neuron or Boss Nucleus, why and how does it *seem* to us that we are unified selves, and, come to think of it, who is this *us* that it so seems to in the first place?

It is almost impossible to talk about human cognition without invoking this "I" or "we" that is the witness and appreciator of the cognitive products, doer of the cognitive deeds. In his recent, and excellent, book, *The Illusion of Conscious Will*, Daniel Wegner lets himself put it this way: "We can't possibly know (let alone keep track of) the tremendous number of mechanical influences on our behavior because we inhabit an extraordinarily complicated machine" (Wegner, 2002, p. 27). These machines "we inhabit" simplify things for *our* benefit. Who or what is this "we" that inhabits the brain? If Hamilton is to be believed, it is a commentator and interpreter with limited access to the actual machinery, more along the lines of an ambassador or press secretary than a president or boss. And this imagery leads straight to Benjamin Libet's vision of "conscious will" as being almost out of the loop. I will briefly describe Libet's experiment, exposing the misinterpretations and explaining its significance for human responsibility. As I put it in *Consciousness Explained,*

> We are not quite "out of the loop" (as they say in the White House), but since our access to information is thus delayed, the most we can do is intervene with

> last-moment "vetoes" or "triggers." Downstream from (unconscious) Command Headquarters, I take no real initiative, am never in on the birth of a project, but do exercise a modicum of executive modulation of the formulated policies streaming through my office. (Dennett, 1991, p. 164)

But I was expressing this view in order to demonstrate its falsehood. I went on to say: "This picture is compelling but incoherent." Others, however, don't see this incoherence. As the sophisticated neuroscientist Michael Gazzaniga has put it: "Libet determined that brain potentials are firing three hundred and fifty milliseconds before you have the conscious intention to act. So before you are aware that you're thinking about moving your arm, your brain is at work preparing to make that movement!" (Gazzaniga, 1998, p.73)

William Calvin, another fine neuroscientist puts it this way:

> My fellow neurophysiologist, Ben Libet has, to everyone's consternation, shown that the brain activity associated with the preparation for movement (something called the "readiness potential")….starts a quarter of a second before you report having decided to move. You just weren't yet conscious of your decision to move, but it was indeed under way…. (Calvin, 1990, p. 80–81)

and Libet himself has recently summarized his own interpretation of the phenomenon thus:

> The initiation of the freely voluntary act appears to begin in the brain unconsciously, well before the person consciously knows he wants to act! Is there, then, any role for conscious will in the performance of a voluntary act? (see Libet, 1985) To answer this it must be recognized that conscious will (W) does appear about 150 msec before the muscle is activated, even though it follows the onset of the RP [readiness potential]. An interval of 150 msec would allow enough time in which the conscious function might affect the final outcome of the volitional process. (Actually, only 100 msec is available for any such effect. The final 50 msec before the muscle is activated is the time for the primary motor cortex to activate the spinal motor nerve cells. During this time the act goes to completion with no possibility of stopping it by the rest of the cerebral cortex.) (Libet, 1999, p. 49)

Only a tenth of a second—100 msec—in which to issue presidential vetoes. As the astute neuroscientist Vilayanur Ramachandran once quipped, "This suggests that our conscious minds may not have free will, but rather 'free won't'!" (Ramachandran, 1998, p. 35) I hate to look a gift horse in the mouth, but I certainly want more free will than that. Can we find any flaws in the reasoning that has led this distinguished group of neuroscientists to this dire conclusion?

My suggestion builds on what I have come to regard as the most important sentence in my 1984 book, *Elbow Room* (p. 143): "(If you make yourself really small, you can externalize virtually everything.)" Stupidly, I made the mistake of putting this sentence in parentheses! Here is what I mean by it. Consider Libet's experiment more closely. When do *you* consciously decide?

Libet needs to hear from you, and he gives you a clock to watch. The experiment depends, in effect, on plotting the intersection of two trajectories:

the visual perception of the clockface and the rising-to-consciousness of the intention or decision or desire to flick the wrist. Putting his two sources of data—your subsequent report and his timed clock positions and EEG tracings—together, he gets an ominous result, but one that depends on his interpretation of the situation, which is only one of several possibilities. Consider three of these alternatives:

A. You are busy making your free decision in the *faculty of practical reasoning* (where all free decisions are made), and you have to wait there for visual contents to be sent over from the *vision center*. How long does this take? If time-pressure is not critical, perhaps the visual content is sent very slowly and is seriously out of date by the time it arrives, like yesterday's newspaper.

B. You are busy watching the clock in the *vision center*, and have to wait for the *faculty of practical reasoning* to send you the results of its latest decision-making. How long does this take? This might be another dawdling transmission, mightn't it?

C. You are sitting where you always sit: in *command headquarters* (otherwise known as the Cartesian Theater), and have to wait for both the *vision center* and the *faculty of practical reasoning* to send their respective outputs to this place, where everything comes together and consciousness happens. If one of these outposts is farther away, or transmits at a slower rate, you will be subject to illusions of simultaneity—if you judge simultaneity by actual arrival time at command headquarters, instead of relying on something like postmarks or time stamps.

Putting the matter this baldly helps, I hope, to see the problems with Libet's picture. What is the presumed implication of these different hypotheses? What would it mean for you to be in one of these places rather than the other? The governing idea is presumably that you can only *act* where you *are*, so if you are not *in* the faculty of practical reasoning when a decision is made there, *you* didn't make it. At best you delegated it. ("I want to be *in* the faculty of practical reasoning. After all, if I'm not *there* when decisions are made, the decisions won't be mine. They will be *its!*") But when you are there, you may get so engrossed in making your decision that "your eyes glaze over" and the vision center's good work goes unattended, never getting to you at all. So, perhaps, you should move back and forth between the faculty of practical reasoning and the vision center. But if that is what you do then it is quite possible that you were in fact conscious of the decision to flick *at the very moment you made it*, but it then took you more than 300 milliseconds to move to the vision center and pick up an image of the clock face showing the position of the moving millisecond mark, so you misjudged the simultaneity because you

lost track of how long it took you to get from place to place. Whew! This is one hypothesis, call it *Strolling You*, that could save free will, by showing that the gap was an illusion, after all. According to this hypothesis, you *consciously* decided to flick when that part of your brain decided to flick (hey, you were *there, at the time,* riding the readiness potential as it was created), but you later misjudged the objective clock time of that decision because of the time it took you to get to the vision center and pick up the latest clockface position.

If you don't like that hypothesis, here is another one that could do the trick, based on alternative C, in which both the vision center and the faculty of practical reasoning are moved out of command headquarters. Call it *Out-of-touch You*. You have outsourced all these tasks, as today's business world would put it, delegating them to subcontractors, but you do keep limited control of their activities from your seat in command headquarters by sending them orders and getting results from them, in a continuous cycle of commands and responses. If asked to think of a reason not to dine out tonight, you send out to your faculty of practical reasoning for a reason, and pretty quick it sends two back: *I'm too tired* and *there's food in the fridge that will spoil if we don't eat it tonight*. How did the faculty come up with these? Why in this order? What operations did it execute to generate them? You haven't a clue—you just know what you sent out for, and recognize that what arrived back is a satisfactory fulfillment of your request. If asked what time it is, you send the appropriate command to the vision center, and it sends back the latest view of the watch on your wrist, with a little help from the *wrist-motion-control center,* but you have no insight into how that collaborative effort was achieved either. Given the problem of variable time delays, you institute a time-stamp system, which works well for most purposes, but you misuse it in Libet's rather unnatural setting. When asked, from your *underprivileged* position in command headquarters to judge just when, exactly, your faculty of practical reasoning issued its flick order (a judgment you are to render in terms of the time-stamps you discern on the streams of reports coming in from both the faculty of practical reasoning and the vision center), you match up the wrong reports. Since you're relying on second-hand information (reports from the two outlying subcontractors), you can easily just be wrong about which event happened first, or whether any two were simultaneous. If you don't like that hypothesis**,** there are others that could be considered, including, of course, all manner of hypotheses that *don't* "save free will" because they tend to confirm Libet's view of the matter: that in the normal course of moral decision-making, *you* in fact have at most 100 milliseconds in which to veto or otherwise adjust decisions made earlier (and elsewhere) unconsciously. Can't we just dismiss the whole sorry lot of them, on the grounds that these hypotheses are wildly unrealistic oversimplifications of what is known about how decision-making works in the brain? Yes indeed, we could, and we should. But when we do that, we don't just dismiss all these fanciful hypotheses that could "save free will" in the face of Libet's data; we must also dismiss Libet's own

hypothesis and all the others that purport to show we only have "free won't." His hypothesis, just as much as those I've just sketched, depends on taking seriously the idea that *you* are restricted to the materials you can get access to from a particular subregion of the brain. How so? Consider his idea of a strictly limited window of opportunity to veto. Libet tacitly presupposes that *you* can't start thinking seriously about whether to veto something until you're conscious of what it is that you might want to veto, and you have to wait 300 milliseconds or more for this, which gives you only 100 milliseconds in which to "act": "This provides a period during which the conscious function could potentially determine whether the volitional process will go on to completion." (Libet, 1993 p. 134.) The "conscious function" waits, in the Cartesian Theater, until the information arrives, and only then *for the first time* has access to it, and can start thinking about what to do about it, whether to veto it, etc. But why couldn't *you* have been thinking ("unconsciously") about whether to veto *Flick!* ever since *you* decided ("unconsciously") to flick, half a second ago? Libet must be assuming that the brain is talented enough to work out the details of implementation on how to flick over that period of time, but only a "conscious function" is talented enough to work on the pros and cons of a veto decision.

In fact, at one point Libet sees this problem and addresses it candidly: "The possibility is not excluded that factors, on which the decision to veto (control) is *based*, do develop by unconscious processes that precede the veto." (1999, p. 51). But if that possibility is not excluded, then the conclusion Libet and others should draw is that the 300-millisecond "gap" has *not* been demonstrated at all. After all, we know that in normal circumstances the brain begins its discriminative and evaluative work as soon as stimuli are received, and works on many concurrent projects at once, enabling us to respond intelligently just in time for many deadlines, without having to stack them up in a queue waiting to get through the turnstile of consciousness before evaluation begins. Libet's 100-millisecond veto window is an artifact of the Cartesian Theater.

When you distribute the work in time and space, you distribute the responsibility in time and space. You are not out of the loop. You *are* the loop.

But why, then, does Libet's Cartesian vision *seem so compelling*? I think we can understand the illusion better if we think about the phenomenon from an evolutionary perspective. For Descartes, the mind was perfectly transparent to itself, with nothing happening out of view, and it has taken more than a century of psychological theorizing and experimentation to erode this ideal of perfect introspectability, which we can now see gets the situation almost backwards. Consciousness of the springs of action is the exception, not the rule, and it requires some rather remarkable circumstances to have evolved at all.

In most of the species that have ever lived, "mental" causation has no need for, and hence does not evolve, any elaborate capacity for self-monitoring. In

general, causes work just fine in the dark, without needing to be observed by anybody, and that is as true of causes in animals' brains as anywhere else. So however "cognitive" an animal's faculties of discrimination might be, the capacity of their outputs to cause the selection of appropriate behavior does not need to be experienced *by anything or anybody.* A bundle of situation-action links of indefinite sophistication can reside in the nervous system of a simple creature and serve its many needs without any further supervision. Its individual actions may need to be guided by a certain amount of internal self-monitoring (specific to the action), in order to make sure, for example, that each predatory swipe snags its target, or to get the berries into the mouth, or to guide the delicate docking with the sexual parts of a conspecific of the opposite sex, but these feedback loops can be as isolated, as local, as the controls that spur the immune system into action when infection looms, or adjust the heart rate and breathing during exercise. (This is the truth behind the deeply misleading intuition that invertebrates, if not "higher, warm-blooded" animals, might be "robots" or "zombies" altogether lacking minds.)

As creatures acquire more and more such behavioral options, however, their worlds become cluttered, and the virtue of tidiness can come to be "appreciated" by natural selection. Many creatures have evolved simple instinctual behaviors for what might be called home improvement, preparing paths, lookouts, hideouts, and other features of their neighborhoods, generally making the local environment easier to get around in, easier to understand. Similarly, when the need arises, creatures evolve instincts for sprucing up their most intimate environments: their own brains, creating paths and landmarks for later use.

The goal unconsciously followed in these preparations is for the creature to come to know its way around itself, and how much of this internal home improvement is accomplished by individual self-manipulation and how much is incorporated genetically is an open empirical question. Along one of these paths, or many of them, lie the innovations that lead to creatures capable of considering different courses of action in advance of committing to any one of them, and weighing them on the basis of some projection of the probable outcome of each.

In the quest by brains to produce a useful future, this is a major improvement over the risky business of *blind* trial and error, since, as Karl Popper once put it, it permits some of your hypotheses to die in your stead. Such Popperian creatures, as I have called them, get to test some of their hunches in informed simulations, rather than risking them in the real world, but they needn't understand the rationale of this improvement in order to reap the benefits. The appreciation of the likely effects of particular actions is built into any such assessment, but the appreciation of the effects of the contemplation itself is a still higher, even more optional, level of self-monitoring. You don't have to know you're a Popperian creature to be one. After all, any chess playing computer considers and discards thousands or millions of possible moves

on the basis of their probable outcomes, and it is manifestly not a conscious or self-conscious agent. (Not yet—the future may hold conscious and even self-conscious robots, which are certainly not impossible.)

What was it that arose in the world to encourage the evolution of a *less unwitting* implementation of Popperian behavioral control? What new environmental complexity favored the innovations in control structure that made this possible? In a word, communication. It is only once a creature begins to develop the activity of communication, and in particular the communication of its actions and plans, that it has to have some capacity for monitoring not just the results of its actions, but of its prior evaluations and formation of intentions as well (McFarland, 1989). At that point, it needs a level of self-monitoring that keeps track of which situation–action schemes are in the queue for execution, or in current competition for execution, and which candidates are under consideration in the faculty of practical reasoning, if that is not too grand a term for the arena in which the competition ensues. How could this new talent arise? We can tell a Just So Story that highlights the key features.

Compare the situation confronting our ancestors (and Mother Nature) to the situation confronting the software engineers who wanted to make computers more user-friendly. Computers are fiendishly complex machines, most of the details of which are nauseatingly convoluted and, for most purposes, beneath notice. Computer-users don't need information on the states of all the flip-flops, the actual location of their data on the disk, and so forth, so software designers created a series of simplifications—even benign distortions in many cases—of the messy truth, cunningly crafted to mesh with, and enhance, the users' pre-existing powers of perception and action. Click and drag, sound effects, icons on desktops, are the most obvious and famous of these, but anybody who cares to dig deeper will find a bounty of further metaphors that help make sense of what is going on inside, but always paying the cost of simplification. As people interacted more and more with computers, they devised a host of new tricks, projects, goals, ways of using and abusing the competences designed for them by the engineers, who thereupon went back to the drawing board to devise further refinements and improvements, which were then used and abused in turn, a co-evolutionary process that continues apace today. The user interface we interact with today was unimagined when computers first appeared, and it is the tip of an iceberg in several senses: not only are the details of what goes on inside your computer hidden; but so are the details of the history of R & D, the false starts, the bad ideas that fizzled before ever reaching the public (as well as the notorious ones that did, and failed to catch on).

A similar process of R & D created the user interface between talking people and other talking people, and it uncovered similar design principles and (free-floating) rationales. It too was co-evolutionary, with people's behaviors, attitudes and purposes evolving in response to the new powers they discovered. Now people could *do things with words* that they could never do before, and the beauty of the whole development was that it *tended* to make those fea-

tures of their complicated neighbors that they were most interested in adjusting readily accessible to adjustment from outside—even by somebody who knew nothing about the internal control system, the brain. These ancestors of ours discovered whole generative classes of behaviors for adjusting the behavior of others, and for monitoring and modulating (and if need be, resisting) the reciprocal adjustment of their own behavioral controls by those others.

The centerpiece metaphor of this co-evolved human user-illusion is the Self, which appears to reside in a place in the brain, the Cartesian Theater, providing a limited, metaphorical outlook on what's going on in our brains. It provides this outlook to others, *and to ourselves*. In fact, we wouldn't exist, as Selves "inhabiting complicated machinery" as Wegner vividly puts it, if it weren't for the evolution of social interactions requiring each human animal to create within itself a subsystem designed for interacting with others. Once created, it could also interact with itself at different times. Until we human beings came along, no agent on the planet enjoyed the curious *non*-obliviousness we have to the causal links that emerged as salient once we human beings began to talk about what we were up to.[a] As Wegner puts it, "People become what they think they are, or what they find that others think they are, in a process of negotiation that snowballs constantly." (p. 314)

When psychologists and neuroscientists devise a new experimental setup or paradigm in which to test nonhuman subjects such as rats or cats or monkeys or dolphins, they often have to devote dozens or even hundreds of hours to training each subject on the new tasks. A monkey, for instance, can be trained to look to the left if it sees a grating moving up and look to the right if it sees a grating moving down. A dolphin can be trained to retrieve an object that looks like (or *sounds* like, to its echolocating system) an object displayed to it by a trainer. All this training takes time and patience, on the part of both trainer and subject. Human subjects in such experiments, however, can usually just be told what is desired of them. After a brief question-and-answer session and a few minutes of practice, we human subjects will typically be as competent in the new environment as any agent ever could be. Of course, we do have to *understand* the representations presented to us in these briefings, and what is asked of us has to be composed of action-parts that fall within the range of things we can do. That is what Wegner means when he identifies voluntary actions as things we can do when asked. If asked to lower your blood pressure or adjust your heartbeat or wiggle your ears, you will not be so ready to comply, though with training not unlike that given to laboratory animals, you may eventually be able to add such feats to your repertoire of voluntary actions.

When language came into existence, it brought into existence the kind of mind that can transform itself on a moment's notice into a somewhat different virtual machine, taking on new projects, following new rules, adopting new

[a]Philosophers may want to compare my Just So Story to Wilfrid Sellars' (1963) myth of "our Rylean ancestors," and Jones, the inventor of "thoughts." My debt to Sellars should be clear to them.

policies. We are transformers. That's what a mind is, as contrasted with a mere brain: the control system of a chameleonic transformer. A virtual machine for making more virtual machines. Non-human animals can engage in voluntary action of sorts. The bird that flies wherever it wants is voluntarily wheeling this way and that, voluntarily moving its wings, and it does this without benefit of language. The distinction embodied in anatomy between what it can do voluntarily (by moving its striated muscles) and what happens autonomically, moved by smooth muscle and controlled by the autonomic nervous system, is not at issue. We have added a layer on top of the bird's (and the ape's and the dolphin's) capacity to decide what to do next. It is not an anatomical layer in the brain, but a functional layer, a virtual layer composed somehow in the micro-details of the brain's anatomy: We can ask each other to do things, and we can ask ourselves to do things. And at least sometimes we readily comply with these requests. Yes, your dog can be "asked" to do a variety of voluntary things, but it can't ask why you make these requests. A male baboon can "ask" a nearby female for some grooming, but neither of them can discuss the likely outcome of compliance with this request, which might have serious consequences for both of them, especially if the male is not the alpha male of the troop. We human beings not only can do things when requested to do them; we can answer inquiries about what we are doing and why. We can engage in the practice of asking, and giving, reasons.

It is this kind of asking, which we can also direct to ourselves, that creates the special category of voluntary actions that sets us apart. Other, simpler intentional systems act in ways that are crisply predictable on the basis of the beliefs and desires we attribute to them based on our surveys of their needs and their history, their perceptual and behavioral talents, but some of our actions are different, in a *morally relevant* way: they result from decisions we make in the course of trying to make sense of ourselves and our own lives (Coleman, 2001).

Once we begin talking about what we're doing, we need to keep track of what we're doing so we can have ready answers to these inquiries. Language requires us to keep track, but also helps us keep track, by helping us categorize and (over)simplify our *agendas*. We cannot help but become amateur auto-psychologists. Nicholas Humphrey and others have spoken of apes and other highly social species as *natural psychologists*, because of the manifest skill and attention they devote to interpreting each other's behavior, but since, unlike academic psychologists—and other human beings—apes never get to compare notes, to argue about attributions of motives and beliefs, their competence as psychologists never obliges them to use explicit representations. With us, it is different. We need to have something to say when asked what the heck we think we're doing. And when we answer, our authority is problematic, as Hamilton noted.

Wegner is right, then, to identify the Self that emerges in his and Libet's experiments as a sort of public-relations agent, a spokesperson instead of a

boss, but these are extreme cases set up to isolate factors that are normally integrated, and we need not identify *ourselves* so closely with such a temporarily isolated self. ("If you make yourself really small, you can externalize virtually everything." (Dennett, 1984, p. 143) Wegner draws our attention to the times—not infrequent among those of us who are "absent-minded"—when we find ourselves with a perfectly conscious thought that just baffles us; it is, as he wonderfully puts it, *conscious but not accessible* (p. 163). (Now why am I standing in the kitchen in front of the cupboard? I know I'm in the place I meant to be, but what did I come in here to get?) At such a moment, *I* have lost track of the context, and hence the *raison d'être*, of this very thought, this conscious experience, and so its meaning (and that's what is most important) is temporarily no more accessible to *me*—the larger me that does the policy-making—than it would be to any third party, any "outside" observer who came upon it. In fact some onlooker might well be able to remind me of what it was I was up to. My capacity to be reminded (re-minded) is crucial, since it is only this that could convince me that this onlooker was right, that this was something *I* was doing. If the thought or project is anyone's, it is mine—it belongs to the me who set it in motion and provided the context in which this thought makes sense; it is just that the part of me that is baffled is temporarily unable to gain access to the other part of me that is the author of this thought.

I might say, in apology, that I was *not myself* when I made that mistake, or forgot what I was about, but this is not the severe disruption of self-control that is observed in schizophrenia, in which the patient's own thoughts are interpreted as alien voices. This is just the fleeting loss of contact that can disrupt a perfectly good plan. A lot of what *you* are, a lot of what you are doing and know about, springs from structures down there in the engine room, causing the action to happen. If a thought of yours is *only* conscious, but not also accessible *to that machinery* (to some of it, to the machinery that needs it), then *you* can't do anything with it, and are left just silently mouthing the damn phrase to yourself, your isolated self, over and over. Isolated consciousness can indeed do nothing much on its own. Nor can it be responsible.

As Wegner notes, "If people will often forget tasks for the simple reason that the tasks have been completed, this signals a *loss of contact* [emphasis added–DCD] with their initial intentions once actions are over—and thus a susceptibility to revised intentions." (p. 167) A loss of contact between what and what? Between a Cartesian Self that "does nothing" and a brain that makes all the decisions? No. A loss of contact between the you that was in charge then and the you that is in charge now. A *person* has to be able to keep in contact with past and anticipated intentions, and one of the main roles of the brain's user-illusion of itself, which I call the self as a center of narrative gravity, is to provide *me* with a means of interfacing with myself at other times. As Wegner puts it, "Conscious will is particularly useful, then, as a guide to ourselves."(p. 328) The perspectival trick we need in order to escape

the clutches of the Cartesian Theater is coming to see that *I,* the larger, temporally and spatially extended self, can control, to some degree, what goes on inside of the simplification barrier, where the decision-making happens, and that is why, as Wegner says, "Illusory or not, conscious will is the person's guide to his or her own moral responsibility for action." (p. 341)

ACKNOWLEDGMENTS

This essay is adapted from a chapter in my forthcoming book, *Freedom Evolves* (Viking Penguin, 2003). It also draws on material in my Nicod Lectures and Daewoo Lectures, forthcoming from MIT Press.

REFERENCES

AINSLIE, GEORGE (2001). *The breakdown of will*. Cambridge: Cambridge University Press.

CALVIN, WILLIAM (1990). *The cerebral symphony: Seashore reflections on the structure of consciousness*. New York: Bantam.

COLEMAN, MARY (2001). *Decisions in action: Reasons, motivation, and the connection between them.* Ph.D. dissertation, Philosophy Department, Harvard University.

DENNETT, DANIEL C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: Bradford Books/MIT Press and Oxford University Press

DENNETT, DANIEL C. (1991). *Consciousness explained*. Boston: Little, Brown

GAZZANIGA, MICHAEL. (1998) *The mind's past*. Berkeley and Los Angeles: University of California Press.

HAMILTON, WILLIAM D. (1996). *Narrow roads of gene land: Vol 1: Evolution of social behaviour.* Oxford: Freeman.

LIBET, BENJAMIN, *et al.* (1983). Time of conscious intention to act in relation to onset of cerebral activities (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*, *106,* 623–642.

LIBET, BENJAMIN (1999). Do we have free will? in Benjamin Libet, Anthony Freeman, and Keith Sutherland (Eds.), *vide infra*, pp. 45–55.

LIBET, BENJAMIN, FREEMAN, ANTHONY & SUTHERLAND, KEITH (1999). *The volitional brain: Towards a neuroscience of free will*. Thorverton, UK: Imprint Academic.

RAMACHANDRAN, V. Quoted in *New Scientist ,* 5 September 1998, p. 35.

SELLARS, WILFRID (1963). Empiricism and the philosophy of mind, in Sellars, *Science, perception and reality.* London: Routledge & Kegan Paul, pp. 127–196.

WEGNER, DANIEL (2002). *The illusion of conscious will*. Cambridge, MA: Bradford Books/MIT Press.