The Self as a Center of Narrative Gravity*

Daniel Dennett

What is a self? I will try to answer this question by developing an analogy with something much simpler, something which is nowhere near as puzzling as a self, but has some properties in common with selves.

What I have in mind is *the center of gravity* of an object.

This is a well-behaved concept in Newtonian physics. But a center of gravity is not an atom or a subatomic particle or any other physical item in the world. It has no mass; it has no color; it has no physical properties at all, except for spatio-temporal location. It is a fine example of what Hans Reichenbach would call an *abstractum*. It is a purely abstract object. It is, if you like , a theorist's fiction. It is not one of the real things in the universe in addition to the atoms. But it is a fiction that has nicely defined, well delineated and well behaved role within physics.

Let me remind you how robust and familiar the idea of a center of gravity is. Consider a chair. Like all other physical objects, it has a center of gravity. If you start tipping it, you can tell more or less accurately whether it would start to fall over or fall back in place if you let go of it. We're all quite good at making predictions involving centers of gravity and devising explanations about when and why things fall over. Place a book on the chair. It, too, has a center of gravity. If you start to push it over the edge, we know that at some point will fall. It will fall when its center of gravity is no longer directly over a point of its supporting base (the chair seat). Notice that that statement is itself virtually tautological. The key terms in it are all interdefinable. And yet it can also figure in explanations that appear to be causal explanations of some sort. We ask "Why doesn't that lamp tip over?" We reply "Because its center of gravity is so low." Is this a causal explanation? It can compete with explanations that are clearly causal, such as: "Because it's nailed to the table," and "Because it's supported by wires."

We can manipulate centers of gravity. For instance, I change the center of gravity of a water pitcher easily, by pouring some of the water out. So, although a center of gravity is a purely abstract object, it has a spatio-temporal career, which I can affect by my actions. It has a history, but its history can include some rather strange episodes. Although it moves around in space and time, its motion can be discontinuous. For instance, if I were to take a piece of bubble gum and suddenly stick it on the pitcher's handle, that would shift the pitcher's center of gravity from point A to point B. But the center of gravity would not have to move through all

* Originally published in in F. Kessel, P. Cole and D. Johnson, eds, *Self and Consciousness: Multiple Perspectives*, Hillsdale, NJ: Erlbaum, 1992.

the intervening positions. As an abstractum, it is not bound by all the constraints of physical travel.

Consider the center of gravity of a slightly more complicated object. Suppose we wanted to keep track of the career of the center of gravity of some complex machine with lots of turning gears and camshafts and reciprocating rods--the engine of a steam-powered unicycle, perhaps. And suppose our theory of the machine's operation permitted us to plot the complicated trajectory of the center of gravity precisely. And suppose--most improbably--that we discovered that in this particular machine the trajectory of the center of gravity was precisely the same as the trajectory of a particular iron atom in the crankshaft. Even if this were discovered, we would be wrong even to *entertain* the hypothesis that the machine's center of gravity was (identical with) that iron atom. That would be a category mistake. A center of gravity is *just* an abstractum. It's just a fictional object. But when I say it's a fictional object, I do not mean to disparage it; it's a wonderful fictional object, and it has a perfectly legitimate place within serious, sober, *echt* physical science.

A self is also an abstract object, a theorist's fiction. The theory is not particle physics but what we might call a branch of people-physics; it is more soberly known as a phenomenology or hermeneutics, or soul-science (*Geisteswissenschaft*). The physicist does an *interpretation*, if you like, of the chair and its behavior, and comes up with the theoretical abstraction of a center of gravity, which is then very useful in characterizing the behaviour of the chair in the future, under a wide variety of conditions. The hermeneuticist or phenomenologist--or anthropologist--sees some rather more complicated things moving about in the world--human beings and animals--and is faced with a similar problem of interpretation. It turns out to be theoretically perspicuous to organize the interpretation around a central abstraction: each person has a *self* (in addition to a center of gravity). In fact we have to posit selves for *ourselves* as well. The theoretical problem of self-interpretation is at least as difficult and important as the problem of other-interpretation.

Now how does a self differ from a center of gravity? It is a much more complicated concept. I will try to elucidate it via an analogy with another sort of fictional object: fictional characters in literature. Pick up *Moby Dick* and open it up to page one. It says, "Call me Ishmael." Call whom Ishmael? Call Melville Ishmael? No. Call Ishmael Ishmael. Melville has created a fictional character named Ishmael. As you read the book you learn about Ishmael, about his life, about his beliefs and desires, his acts and attitudes. You learn a lot more about Ishmael then Melville ever explicitly tells you. Some of it you can read in by implication. Some of it you can read in by extrapolation. But beyond the limits of such extrapolation fictional worlds are simply *indeterminate*. Thus, consider the following question (borrowed from David Lewis's "Truth and Fiction," *American Philosophical Quarterly*, 1978, **15**, pp.37-46). Did Sherlock Holmes have three nostrils? The answer of course is no, but not because Conan Doyle ever says that he doesn't, or

that he has two, but because we're entitled to make that extrapolation. In the absence of evidence to the contrary, Sherlock Holmes' nose can be supposed to be normal. Another question: Did Sherlock Holmes have a mole on his left shoulder blade? The answer to this question is neither yes nor no. Nothing about the text or about the principles of extrapolation from the text permit an answer to that question. There is simply no fact of the matter. Why? Because Sherlock Holmes is a merely fictional character, created by, or constituted out of, the text and the culture in which that text resides.

This indeterminacy is a fundamental property of fictional objects which strongly distinguishes them from another sort of object scientists talk about: theoretical entities, oor what Reichenbach called *illata*--inferred entities, such as atoms, molecules and neutrinos. A logician might say that the "rpinciple of bivalence" does not hold for fictional objects. That is to say, with regard to any actual man, living or dead, the question of whether or not he has or had a mole on his left shoulder blade has an anwer, yes or no. Did Aristotle has such a mole? There is a fact of the matter even if we can never discover it. But with regard to a fictional character, that question may have no answer at all.

We can imagine someone, a benighted literary critic, perhaps, who doesn't understand that fiction is fiction. This critic has a strange theory about how fiction works. He thinks that something literally magical happens when a novelist writes a novel. When a novelist sets down words on paper, this critic says (one often hears claims like this, but not meant to be taken completely literally), the novelist actually *creates a world*. A litmus test for this bizarre view is the principle of bivalence: when our imagined critic speaks of a fictional world he means a strange sort of *real* world, a world in which the principle of bivalence holds. Such a critic might seriously wonder whether Dr Watson was *really* Moriarty's second cousin, or whether the conductor of the train that took Holmes and Watson to Aldershot was also the conductor of the train that brought them back to London. That sort of question can't properly arise if you understand fiction correctly, of course. Whereas analogous questions about historical personages have to have yes or no answers, even if we may never be able to dredge them up.

Centers of gravity, as a fictional objects, exhibit the same feature. They have only the properties that the theory that constitutes them endowed them with. If you scratch your head and say, "I wonder if maybe centers of gravity are really neutrinos!" you have misunderstood the theoretical status of a center of gravity.

Now how can I make the claim that a self--your own real self, for instance--is rather like a fictional character? Aren't all *fictional* selves dependent for their very creation on the existence of *real* selves? It may seem so, but I will argue that this is an illusion. Let's go back to Ishmael. Ishmael is a fictional character, although we can certainly learn all about him. One might find him in many regards more real than many of one's friencds. But, one thinks, Ishmael was created by Melville, and Melville is a real character--was a real character. A real self. Doesn't this show that

it takes a real self to create a fictional self? I think not, but If I am to convince you, I must push you through an exercise of the imagination.

First of all, I want to imagine something some of you may think incredible: a novel-writing machine. We can suppose it is a product of artificial intelligence research, a computer that has been designed or programmed to write novels. But it has not been designed to write any particular novel. We can suppose (if it helps) that it has been given a great stock of whatever information it might need, ans some partially random and hence unpredictable ways of starting the seed of a story going, and building upon it. Now imagine that the designers are sitting back, wondering what kind of novel their creation is going to write. They turn the thing on and after a while the high speed printer begins to go clickety-clack and out comes the first sentence. "Call me Gilbert," it says. What follows is the apparent autobiography of some fictional Gilbert. Now Gilbert is a fictional, created self but its creator is no self. Of course there were human designers who designed the machine, but they didn't design Gilbert. Gilbert is a product of a design or invention process in which there aren't any selves at all. That is, I am *stipulating* that this is not a conscious machine, not a "thinker." It is a dumb machine, but it does have the power to write a passable novel. (IF you think this is striclty impossible I can only challenge you to show why you think this must be so, and invite you read on; in the end you may not have an interest in defending such a precarious impossibilility-claim.)

So we are to imagine that a passable story is emitted from the machine. Notice that we can perform the same sort of literary exegesis with regard to this novel as we can with any other. In fact if you were to pick up a novel at random out of a library, you could not tell with certainty that it wasn't written by something like this machine. (And if you're a New Critic you shouldn't care.) You've got a text and you can interpret it, and so you can learn the story, the life and adventures of Gilbert. Your expectations and predictions, as you read, and your interpretive reconstruction of what you have already read, will congeal around the central node of the fictional character, Gilbert.

But now I want to twiddle the knobs on this thought experiment. So far we've imagined the novel, *The Life and Times of Gilbert*, clanking out of a computer that is just a box, sitting in the corner of some lab. But now I want to change the story a little bit and suppose that the computer has arms and legs--or better: wheels. (I don't want to make it too anthropomorphic.) It has a television eye, and it moves around in the world. It also begins its tale with "Call me Gilbert," and tells a novel, but now we notice that if we do the trick that the New Critics say you should never do, and *look outside the text*, we discover that there's a truth-preserving interpretation of that text in the real world. The adventures of Gilbert, the fictional character, now bear a striking and presumably non-coincidental relationship to the adventures of this robot rolling around in the world. If you hit the robot with a baseball bat, very shortly thereafter the story of Gilbert includes his being hit with a baseball bat by somebody who looks like you. Every now and then the robot gets

locked in the closed and then says "Help me!" Help whom? Well, help Gilbert, presumably. But who is Gilbert? Is Gilbert the robot, or merely the fictional self created by the robot? If we go and help the robot out of the closet, it sends us a note: "Thank you. Love, Gilbert." At this point we will be unable to ignore the fact that the fictional career of the fictional Gilbert bears an intersting resemblance to the "career" of this mere robot moving through the world. We can still maintain that the robot's *brain*, the robot's computer, really knows nothing about the world; *it's* not a self. It's just a clanky computer. It doesn't know what it's doing. It doesn't even know that it's creating a fictional character. (The same is just as true of your brain; *it* doesn't know what it's doing either.) Nevertheless, the patterns in the behavior that is being controlled by the computer are interpretable, by us, as accreting biography--telling the narrative of a self. But we are not the only interpreters. The robot novelist is also, of course, an interpreter: a *self*-interpreter, providing its own account of its activities in the world.

I propose that we take this analogy seriously. "Where is the self?" a materialist philosopher or neuroscientist might ask. It is a category mistake to start looking around for the self in the brain. Unlike centers of gravity, whose sole property is their spatio-temporal position, selves have a spatio-temporal position that is only grossly defined. Roughly speaking, in the normal case if there are three human beings sitting on a park bench, there are three selves there, all in a row and roughly equidistant from the fountain they face. Or we might use a rather antique turn of phrase and talk about how many *souls* are located in the park. ("All twenty souls in the starboard lifeboat were saved, but those that remained on deck perished.")

Brain research may permit us to make some more fine-grained localizations, but the capacity to achieve *some* fine-grained localization does not give one grounds for supposing that the process of localization can continue indefinitely and that the day will finally come when we can say, "That cell there, right in the middle of hippocampus (or wherever)--that's the self!"

There's a big difference, of course, between fictional characters and our own selves. One I would stress is that a fictional character is usually encountered as a *fait accompli*. After the novel has been written and published, you read it. At that point it is too late for the novelist to render determinate anything indeterminate that strikes your curiosity. Dostoevesky is dead; you can't ask him what *else* Raskolnikov thought while he sat in the police station. But novels don't have to be that way. John Updike has written three novels about Rabbit Angstrom: *Rabbit Run*, *Rabbit Redux*, and *Rabbit is Rich*. Suppose that those of us who particularly liked the first novel were to get together and compose a list of questions for Updike--things we wished Updike has talked about in that first novel, when Rabbit was a young former basketball star. We could send our questions to Updike and ask him to consider writing another novel in the series, only this time not continuing the chronological sequence. Like Lawrence Durrell's *Alexandria Quarter*, the Rabbit series could include another novel about Rabbit's early days when he was still playing basketball, and this novel could answer our questions.

Notice what we would *not* be doing in such a case. We would not be saying to Updike, "Tell us the answers that you already know, the answers that are already fixed to those questions. Come on, let us know all those secrets you've been keeping from us." Nor would we be asking Updike to do research, as we might ask the author of a multi-volume biography of a real person, We would be asking him to write a new novel, to invent some more novel for us, on demand. And if he acceded, he would enlarge and make more determinate the character of Rabbit Angstrom in the process of writing the new novel. In this way matters which are indeterminate at one time can become determined later by a creative step.

I propose that this imagined exercise with Updike, getting him to write more novels on demand to answer our questions, is actually a familiar exercise. That is the way we treat each other; that is the way we are. We cannot undo those parts of our pasts that are determinate, but our selves are constantly being made more determinate as we go along in response to the way the world impinges on us. Of course it is also possible for a person to engage in auto-hermeneutics, interpetation of one's self, and in particular to go back and think about one's past, and one's memories, and to rethink them and rewrite them. This process does change the "fictional" character, the character that you are, in much the way that Rabbit Angstrom, after Updike writes the second novel about him as a young man, comes to be a rather different fictional character, determinate in ways he was never determinate before. This would be an utterly mysterious and magical prospect (and hence something no one should take seriously) *if the self were anything but an abstractum*.

I want to bring this out by extracting one more feature from the Updike thought experiment. Updike might take up our request but then he might prove to be forgetful. After all, it's been many years since he wrote *Rabbit Run*. He might not want to go back and reread it carefully; and when he wrote the new novel it might end up being inconsistent with the first. He might have Rabbit being in two places at one time, for instance. If we wanted to settle what the *true* story was, we'd be falling into error; there is no true story. In such a circumstance there would be simply be a failure of coherence of all the data that we had about Rabbit. And because Rabbit is a fictional character, we wouldn't smite our foreheads in wonder and declare "Oh my goodness! There's a rift in the universe; we've found a contradiction in nature!" Nothing is easier than contradiction when you're dealing with fiction; a fictional character can have contradictory properties because it's *just* a fictional character. We find such contradictions intolerable, however, when we are trying to interpret something or someone, even a fictional character, so we typically *bifurcate* the character to resolve the conflict.

Something like this seems to happen to real people on rare occasions. Consider the putatively true case histories recorded in *The Three Faces of Eve* and *Sybil*. (Corbett H. Thigpen and Hervey Cleckly, *The Three Faces of Eve*, McGraw Hill, 1957, and Flora Rheta Schreiber, *Sybil*, Warner paperback, 1973.) Eve's three faces were the faces of three distinct personalities, it seems, and the woman portrayed in

Sybil had *many* different selves, or so it seems. How can we make sense of this? Here is one way--a solemn, skeptical way favored by some of the psychotherapists with whom I've talked about such cases: when Sybil went in to see her therapist the first time, she wasn't several different people rolled into one body. Sybil was a novel-writing machine that fell in with a very ingenious questioner, a very eager reader. And together they collaborated--innocently--to write many, many chapters of a new novel. And, of course, since Sybil was a sort of living novel, she went out and engaged in the world with these new selves, more or less created on demand, under the eager suggestion of a therapist.

I now believe that this is overly skeptical. The population explosion of new characters that typically follows the onset of psychotherapy for sufferers of Multiple Personality Disorder (MPD) is probably to be explained along just these lines, but there is quite compelling evidence in some cases that some multiplicity of selves (two or three or four, let us say) had already begun laying down biography before the therapist came along to do the "reading". And in any event, Sybil is only a strikingly pathological case of something quite normal, a behavior pattern we can find in ourselves. We are all, at times, confabulators, telling and retelling ourselves the story of our own lives, with scant attention to the question of truth. Why, though do we behave this way? Why are we all such inveterate and inventive autobiographical novelists? As Umberto Maturana has (uncontroversially) observed: "Everything said is said by a speaker to another speaker that may be himself." But why should one talk to oneself? Why isn't that an utterly idle activity, as systematically futile as trying to pick oneself up by one's own bootstraps?

A central clue comes from the sort of phenomena uncovered by Michael Gazzaniga's research on those rare individuals--the "split-brain subjects"--whose *corpus callosum* has been surgically severed, creating in them two largely independent cortical hempisheres that can, on occasion, be differently informed about the current scene. Does the operation *split* the self in two? After the operation, patients normally exhibit no signs of psychological splitting, appearing to be no less unified than you or I except under particularly contrived circumstances. But on Gazzaniga's view, this does not so much show that the patients have preserved their pre-surgical unity as that the unity of normal life is an illusion.

According to Gazzaniga, the normal mind is *not* beautifully unified, but rather a problematically yoked-together bundle of partly autonomous systems. All parts of the mind are not equally accessible to each other at all times. These modules or systems sometimes have internal communication problems which they solve by various ingenious and devious routes. If this is true (and I think it is), it may provide us with an answer to a most puzzling question about conscious thought: what good is it? Such a question begs for a evolutionary answer, but it will have to speculative, of course. (It is not critical to my speculative answer, for the moment,

where genetic evolution and transmission breaks off and cultural evolution and transmission takes over.)

In the beginning--according to Julian Jaynes (*The Origins of Consciousness in the Breakdown of the Bicameral Mind*, Boston: Hougton Mifflin, 1976), whose account I am adapting--were speakers, our ancestors, who weren't really conscious. They spoke, but they just sort of blurted things out, more or less the way bees do bee dances, or the way computers talk to each other. That is not conscious communication, surely. When these ancestors had problems, sometimes they would "ask" for help (more or less like Gilbert saying "Help me!" when he was locked in the closet), and sometimes there would be somebody around to hear them. So they got into the habit of asking for assistance and, particularly, asking questions. Whenever they couldn't figure out how to solve some problem, they would ask a question, addressed to no one in particular, and sometimes whoever was standing around could answer them. And they also came to be designed to be provoked on many such occasions into answering questions like that--to the best of their ability--when asked.

Then one day one of our ancestors asked a question in what was apparently an inappropriate cirucmstance: thre was nobody around to be the audience. Strangely enough, he heard his own question, and this stimulated him, cooperatively, to think of an answer, and sure enough the answer came to him. He had established, without realizing what he had done, a communication link between two parts of his brain, between which there was, for some deep biological reason, an accessability problem. One component of the mind had confronted a problem that another component could solve; if only the problem could be posed for the latter component! Thanks to his habit of asking questions, our ancestor stumbled upon a route via the ears. What a discovery! Sometimes talking and listening to yourself can have wonderful effects, not otherwise obtainable. All that is needed to make sense of this idea is the hypothesis that the modules of the mind have different capacities and ways of doing things, and are not perfectly interaccessible. Under such circumstances it could be true that the way to get yourself to figure out a problem is to tickle your ear with it, to get that part of your brain which is best stimulated by *hearing* a question to work on the problem. Then sometimes you will find yourself with the answer you seek on the tip of your tongue.

This would be enough to establish the evolutionary endorsement (which might well be only culturally transmitted) of the behavior of *talking to yourself*. But as many writers have observed, conscious thinking seems--much of it--to be a variety of particularly efficient and private talking to oneself. The evolutionary transition to thought is then easy to conjure up. All we have to suppose is that the route, the circuit that at first went via mouth and ear, got shorter. People "realized" that the actual vocalization and audition was a rather inefficient part of the loop. Besides, if there were other people around who might overhear it, you might give away more information than you wanted. So what developed was a habit of subvocalization, and this in turn could be streamlined into conscious, verbal thought.

In his posthumous book *On Thinking* (ed. Konstantin Kolenda, Totowa New Jersey, Rowman and Littlefield, 1979), Gilbert Ryle asks: "What is *Le Penseur* doing?" For behaviorists like Ryle this is a real problem. One bit of chin-on-fist-with-knitted-brow looks pretty much like another bit, and yet some of it seems to arrive at good answers and some of it doesn't. What can be going on here? Ironically, Ryle, the arch-behaviorist, came up with some very sly suggestions about what might be going on. Conscious thought, Ryle claimed, should be understood on the model of self-teaching, or better, perhaps: self-schooling or training. Ryle had little to say about how this self-schooling might actually work, but we can get some initial understanding of it on the supposition that we are *not* the captains of our ships; there is no conscious self that is unproblematically in command of the mind's resources. Rather, we are somewhat disunified. Our component modules have to act in opportunistic but amazingly resourceful ways to produce a modicum of behavioral unity, *which is then enhanced by an illusion of greater unity*.

What Gazzaniga's research reveals, sometimes in vivid detail, is how this must go on. Consider some of his evidence for the extraordinary resourcefulness exhibited by (something in) the right hemisphere when it is faced with a communication problem. In one group of experiments, split-brain subjects must reach into a closed bag with the left hand to feel an object, which they are then to identify verbally. The sensory nerves in the left hand lead to the right hemisphere, whereas the control of speech is normally in the left hemisphere, but for most of us, this poses no problem. In a normal person, the left hand can know what the right hand is doing thanks to the corpus collosum, which keeps both hemispheres mutually informed. But in a split-brain subject, this unifying link has been removed; the right hemisphere gets the information about the touched object from the left hand, but the left, language-controlling, hemisphere must make the identification public. So the "part which can speak" is kept in the dark, while the "part which knows" cannot make public its knowledge.

There is a devious solution to this problem, however, and split-brain patients have been observed to discover it. Whereas ordinary tactile sensations are represented contralaterally--the signals go to the opposite hemisphere--pain signals are also represented ipsilaterally. That is, thanks to the way the nervous system is wired up, pain stimuli go to both hemispheres. Suppose the object in the bag is a pencil. The right hemisphere will sometimes hit upon a very clever tactic: hold the pencil in your left hand so its point is pressed hard into your palm; this creates pain, and lets the left hemisphere know there's something sharp in the bag, which is enough of a hint so that it can begin guessing; the right hemisphere will signal "getting warmer" and "got it" by smiling or other controllable sings, and in a very short time "the subject"--the *apparently* unified "sole inhabitant" of the body--will be able to announce the correct answer.

Now either the split-brain subjects have developed this extraordinarily devious talent as a reaction to the operation that landed them with such radical

accessibility problem, or the operation *reveals*--but does not create--a virtuoso talent to be found also in normal people. Surely, Gazzaniga claims, the latter hypothesis is the most likely one to investigate. That is, it does seem that we are all virtuoso novelists, who find ourselves engaged in all sorts of behavior, more or less unified, but sometimes disunified, and we always put the best "faces" on it we can. We try to make all of our material cohere into a single good story. And that story is our autobiography.

The chief fictional character at the center of that autobiography is one's *self*. And if you still want to know what the self *really* is, you're making a category mistake. After all, when a human being's behavioral control system becomes seriously impaired, it can turn out that the best hermeneutical story we can tell about that individual says that there is more than one character "inhabiting" that body. This is quite possible on the view of the self that I have been presenting; it does not require any fancy metaphysical miracles. One can discover multiple selves in a person just as unproblematically as one could find Early Young Rabbit and Late Young Rabbit in the imagined Updike novels: all that has to be the case is that the story doesn't cohere around one self, one imaginary point, but coheres (coheres much better, in any case) around two different imaginary points.

We sometimes encounter psychological disorders, or surgically created disunities, where the only way to interpret or make sense of them is to posit in effect two centers of gravity, two selves. One isn't creating or discovering a little bit of ghost stuff in doing that. One is simply creating another abstraction. It is an abstraction one uses as part of a theorical apparastus to understand, and predict, and make sense of, the behavior of some very complicated things. The fact that these abstract selves seem so robust and real is not surprising. They are much more complicated theoretical entities than a center of gravity. And remember that even a center of gravity has a fairly robust presence, once we start playing around with it. But no one has ever seen or ever will see a center of gravity. As David Hume noted, no one has ever seen a self, either.

"For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a percepton, and never can observe anything but the perception.... If anyone, upon serious and unprejudiced reflection, thinks he has a different notion of *himself*, I must confess I can reason no longer with him. All I can allow him is, that he may be in the right as well as I, and that we are essentially different in this particular. He may, perhaps, perceive something simple and continued, which he calls *himself*; though I am certain there is no such principle in me." (*Treatise on Human Nature*, I, IV, sec. 6.)