Self to Self

Selected Essays

J. DAVID VELLEMAN
New York University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521670241

First published 2006
Reprinted 2007

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Velleman, James David.
Self to Self : selected essays / J. David Velleman.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-85429-6 (hardcover) – ISBN 0-521-67024-1 (pbk.)
1. Self. 2. Self (Philosophy) 3. Kant, Immanuel, 1724–1804 – Ethics. I. Title.
BF697.V45 2005
126–dc22 2005008114

ISBN 978-0-521-85429-0 Hardback
ISBN 978-0-521-67024-1 Paperback

Transferred to digital printing 2009

9

The Self as Narrator

Many philosophers have thought that human autonomy includes, or perhaps even consists in, a capacity for self-constitution – a capacity, that is, to define or invent or create oneself.[1] Unfortunately, self-constitution sounds not just magical but paradoxical, as if the rabbit could go solo and pull himself out of the hat. Suspicions about the very idea of this trick have sometimes been allayed by appeal to the political analogy implicit in the term "self-constitution": a person is claimed to constitute himself in the same way as a polity does, by writing, ratifying, and revising articles of constitution.[2] But a polity is constituted, in the first instance, by its

[1] A list of philosophers who have held this view would include Charles Taylor (*Sources of the Self: The Making of the Modern Identity* [Cambridge, MA: Harvard University Press, 1989]; *Human Agency and Language* [Cambridge: Cambridge University Press, 1985]); Harry Frankfurt (*The Importance of What We Care About* [Cambridge: Cambridge University Press, 1987]); Christine Korsgaard (*The Sources of Normativity* [Cambridge: Cambridge University Press, 1996]; "Self-Constitution in the Ethics of Plato and Kant," *Journal of Ethics* 3 [1999]: 1–29); Tamar Schapiro ("What Is a Child?" *Ethics* 109 [1999]: 715–38); and Michael Bratman ("Reflection, Planning, and Temporally Extended Agency," *Philosophical Review* 109 [2000]: 35–61).

[2] See, especially, Schapiro.

The material in this chapter was first presented to a seminar on the self, taught in the fall of 1999 at the University of Michigan. Versions of the chapter have been presented to the philosophy departments of the University of Pittsburgh, the University of Maryland, the University of Chicago, and the University of Göttingen; to a conference on Morality and the Arts at the University of California, Riverside, with John Martin Fischer serving as commentator; and as one of the Jerome Simon Lectures at the University of Toronto. I have received helpful comments from the audiences on these occasions as well as from Linda Brakel and Dan Dennett. The chapter first appeared in *Autonomy and the Challenges to Liberalism: New Essays*, edited by Joel Anderson and John Christman (Cambridge: Cambridge University Press, 2005), 56–57, and is reprinted here with the permission of the publisher.

constituent persons, who are constituted antecedently to it; and suspicions therefore remain about the idea of self-constitution at the level of the individual person.

One philosopher has tried to save personal self-constitution from suspicions of paradox by freely admitting that it is a trick. A real rabbit can't pull himself out of a hat, according to this philosopher, but an illusory rabbit can appear to do so: the secret of the trick is that the rabbit isn't real. We ask, "But if the rabbit isn't real – and there's no magician, either – then who is performing the trick?" He replies, "Why, of course: the hat." A rabbit can't pull himself out of a hat, but a hat can make it appear that a rabbit is pulling himself out of it.

Notwithstanding my frivolous analogy, I think that there is much to be learned from this view of self-constitution, and so I propose to examine it in detail and to offer my own variation on it. The philosopher in question is Daniel Dennett, and his view is that the autonomous person (the rabbit) is an illusion conjured up by the human organism (the hat).[3] In the end, I will adopt most of Dennett's view, except for the part about the rabbit's being unreal. In my view, the rabbit really does pull himself out of the hat, after all.

Dennett's metaphor for this process is not sleight-of-hand but fiction. In Dennett's metaphor, the self is the non-existent author of a merely fictional autobiography composed by the human organism, which neither is nor embodies a real self.[4] So understood, the self has the status of an *abstractum*, a fictional object that we "use as part of a theoretical apparatus to understand, and predict, and make sense of, the behavior of some very complicated things"[5] – namely, human beings, including ourselves.

Dennett compares the human's autobiography to the spider's web or the beaver's dam:

Our fundamental tactic of self-protection, self-control, and self-definition is not spinning webs or building dams, but telling stories, and more particularly

[3] "The Origins of Selves," *Cogito* 3 (1989): 163–73 [hereinafter OS]; "The Reality of Selves," in *Consciousness Explained* (Boston: Little, Brown and Company, 1991), Chapter 13 [RS]; "The Self as a Center of Narrative Gravity," in *Self and Consciousness: Multiple Perspectives*, eds., Frank S. Kessel, Pamela M. Cole, and Dale L. Johnson (Hillsdale, NJ: Erlbaum Associates, 1992), 103–115 [CNG]; with Nicholas Humphrey, "Speaking for Our Selves," reprinted in *Brainchildren: Essays on Designing Minds* (Cambridge, MA: MIT Press, 1998), 31–58 [SO].

[4] Dennett describes his view as a "middle-ground position" on the question "whether there really are selves" (RS, 413).

[5] CNG, 114–15.

concocting and controlling the story we tell others – and ourselves – about who we are. [...] These strings or streams of narrative issue forth *as if* from a single source – not just in the obvious physical sense of flowing from just one mouth, or one pencil or pen, but in a more subtle sense: their effect on any audience is to encourage them to (try to) posit a unified agent whose words they are, about whom they are: in short, to posit a *center of narrative gravity.* [RS, 418]

The point of this last phrase is that an object's physical center of gravity can figure in legitimate scientific explanations but mustn't be identified with any physical part of the object:

That would be a category mistake. A center of gravity is *just* an abstractum. It is just a fictional object. But when I say it is a fictional object, I do not mean to disparage it; it is a wonderful fictional object, and it has a perfectly legitimate place within serious, sober, *echt* physical science. [CNG, 104]

Similarly, the "unified agent" conjured up by our narrative is a theoretical abstraction, but it too has a legitimate place in a serious theory. Dennett concludes the analogy as follows:

[W]e are 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 are making a category mistake. [CNG, 114]

What exactly is the category mistake that we make about the self, according to Dennett? I shall first attempt to identify the mistake, and then I'll consider whether it really is a mistake. Specifically, I'll ask whether Dennett himself can afford to call it a mistake, given the philosophical commitments he undertakes in the course of diagnosing it. I shall argue that in at least some respects, the conception of the self that Dennett calls mistaken is in fact likely to be correct.

In arguing against Dennett's diagnosis of this mistake, I shall not be arguing against his positive conception of the self as the fictive protagonist of a person's autobiography.[6] On the contrary, I'll argue that Dennett's positive conception of the self is largely right. My only disagreement with

[6] I use the term "fictive" because, to my ear, it shares with "fictional" the sense of "invented" or "made up," but not the sense of "untrue." Those who do not already share these linguistic intuitions should take them as stipulated hereby.

Dennett will be that, whereas he regards an autobiography as fictive and consequently false in characterizing its protagonist, I regard it as both fictive and true. We invent ourselves, I shall argue, but we really are the characters whom we invent.

Dennett describes our mistaken conception as "the myth of selves as brain-pearls, particular concrete, countable things rather than abstractions."[7] Sometimes he suggests that this myth mistakenly credits the self with physical existence, as "a proper physical part of an organism or a brain."[8] But he also considers a version of the myth in which the self resides in software rather than hardware, as "a supervisory brain program, a central controller, or whatever."[9] Mostly, Dennett relies on metaphors that can be read as alluding either to hardware or software: the "Oval Office in the brain, housing a Highest Authority"[10] or "the Cartesian Theater with its Witness or Central Meaner"[11] or "the central headquarters responsible for organizing and directing all the subsidiary bureaucracies that keep life and limb together."[12]

Dennett cannot be faulted for describing the self in metaphorical terms. His thesis, after all, is that the self is like one of those mythical beasts that incorporate parts from different creatures and straddle boundaries between different realms, in a way that defies literal description. Yet unless we understand what Dennett thinks is wrong with our conception of the self, we cannot understand what he thinks is right about his own, alternative conception. So we must look behind Dennett's metaphors for the error that they purport to reveal.

In Dennett's view, our error about the self is to assume that the protagonist of a human being's autobiography is identical with the author. Dennett imagines that his own autobiography opens in the manner of *Moby Dick* – "Call me Dan" – and he claims that this opening sentence would prompt us to apply that name to "the theorists' fiction created by . . . well, not by me but by my brain [. . .]."[13] In Dennett's view, then, the author of his autobiography is his brain, whereas the "me" whom we call Dan is a purely fictional narrator, who is no more the real author

[7] RS, 424. See p. 423: "independently existing soul-pearls."
[8] RS, 420.
[9] RS, 420.
[10] RS, 428.
[11] RS, 422.
[12] OS, 163.
[13] RS, 429.

of the story than Ishmael is the author of the story that begins "Call me Ishmael." Dennett concludes:

> Our tales are spun, but for the most part we don't spin them; they spin us. Our human consciousness, and our narrative selfhood, is their product, not their source. [RS, 418]

But in what respect does the real source of Dennett's autobiography differ from the fictional source that it conjures up for itself? Why should Dan be compared to Ishmael rather than the author of a veridical autobiography, who really is identical with the protagonist of his story?

This question is especially pressing in light of the sophistication with which Dennett is obliged to credit his real autobiographer. The brain that composes Dennett's autobiography has to be so clever as to approximate the powers of its supposedly fictional protagonist. We may therefore suspect that Dennett, now in his capacity as philosopher, has tacitly posited the existence of a real self to serve as the inventor of the supposedly fictional one. Dennett anticipates and counters this suspicion:

> 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 us go back to Ishmael. Ishmael is a fictional character [...]. 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. [CNG, 107]

The exercise mentioned here is to imagine a robot that emits a running narration of its life, as the story of a character named Gilbert:

> "Call me Gilbert," it says. What follows is the apparent autobiography of this 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 did not design Gilbert. Gilbert is the product of a process in which there are no selves at all. [*Ibid.*]

Dennett insists that he is not committed to crediting the robot with selfhood:

> 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. [*Ibid.*]
>
> [T]he robot's *brain*, the robot's computer, really knows nothing about the world; *it* is 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 this fictional character. (The same is just as true of your brain: *it* doesn't know what it's doing either.) [CNG, 108]

One might challenge this stipulation as self-contradictory. Stipulating a "dumb machine" that writes a "passable novel," one might think, is like stipulating a blind man who sees. If someone sees, then he isn't really blind; and if something writes a passable novel, then it can't be all that dumb, no matter how loudly it may clank.[14] How, then, can Dennett claim that the computer generating Gilbert's story doesn't know what it's doing?

Part of the answer is that, according to Dennett, the computer isn't conscious; but I want to set aside the concept of consciousness, which is only one aspect of selfhood. To be sure, Gilbert's autobiographer portrays him as conscious, while Dennett denies that he really is. But the robot's claim to be conscious is not quite the same as his claim to be a self. For as we have seen, claiming to be a self entails claiming not only the status of "Witness," who is the subject of experience, but also that of "Central Meaner," "central controller," or "Highest Authority."[15] Indeed, Dennett defines a center of narrative gravity as a fictional "unified agent."[16] Leaving aside the question whether Gilbert's autobiographer is conscious, then, we can ask whether he really is a unified agent in the sense that would satisfy the terms of this fiction.

Here again, one might think that Dennett's stipulation is incoherent, on the grounds that describing something as the author of a novel already entails describing it as a unified agent. Yet I am willing to grant, for the sake of argument, that a passable novel could be authored by a machine endowed with no "Highest Authority," "Central Meaner," or other ironically capitalized locus of agency. What I suggest, however, is that Dennett has equipped Gilbert's and Dan's autobiographers with more than the mere capacity to produce passable novels, and that in doing so, he has implicitly equipped them with enough of a self to be agents.

Dennett denies agency to the inventors of Gilbert and Dan primarily by denying them agential unity. He defends this denial by citing the example

[14] If the objection here is merely that writing a passable novel is an activity that is most perspicuously interpreted as the product of a conscious thinker, then Dennett can of course agree, since he believes that positing a conscious thinker, Gilbert, is the most perspicuous way of interpreting the novel-writing robot. What he denies is that writing a novel requires a real, conscious thinker of the sort that would be postulated by such an interpretation.

[15] Quoted at notes 9–11.

[16] RS, 418, quoted after note 13.

of a termite colony:

> The revisionist case is that there really is no proper-self: none of the fictive-selves – including one's own firsthand version – corresponds to anything that actually exists in one's head.
>
> At first sight this might not seem reasonable. Granted that whatever *is* inside the head might be difficult to observe, and granted that it might also be a mistake to talk about a "ghostly supervisor," nonetheless there surely has to be some kind of a supervisor in there: a supervisory brain program, a central controller, or whatever. How else could anybody function – as most people clearly do function – as a purposeful and relatively well-integrated agent?
>
> The answer that is emerging from both biology and Artificial Intelligence is that complex systems can in fact function in what seems to be a thoroughly "purposeful and integrated" way simply by having *lots of subsystems doing their own thing* without any central supervision. Indeed most systems on earth that appear to have central controllers (and are usefully described as having them) do not. The behavior of a termite colony provides a wonderful example of it. The colony as a whole builds elaborate mounds, gets to know its territory, organizes foraging expeditions, sends out raiding parties against other colonies, and so on. [. . .] Yet, in fact, all this group wisdom results from nothing other than myriads of individual termites, specialized as several different castes, going about their individual business – influenced by each other, but quite uninfluenced by any master-plan. [SO, 39–40][17]

Dennett illustrates the unreality of central supervision in humans with the phenomenon of Multiple Personality Disorder (MPD). Writing with a collaborator, Nicholas Humphrey, he hypothesizes that a child subjected to severe abuse may be forced to invent more than one fictional self, whereupon the child is obliged to elect one of these fictional characters as "Head of Mind," who can then be occasionally deposed by competitors.[18] The currently active personality purports to be in control, but we who observe the succession of pretended controllers know that, in reality, nobody is home.

There is no doubt but that Dennett's fictionalism about the self provides an attractive explanation for the phenomenon diagnosed as MPD. According to Dennett, the self is like an imaginary friend from our childhood – an especially close imaginary friend who became not merely our *alter* ego but, so to speak, our *auto* ego. Just as some of us may have developed more than one imaginary friend, if we had unusual emotional

[17] See also OS, 167–68, and RS, 416, where Dennett remarks, "There is [. . .] no Oval Office in the anthill," just as he subsequently remarks that "there is no Oval Office in the brain" [RS, 429].

[18] SO, 41. For another narrative-based analysis of MPD, see Valerie Gray Hardcastle and Owen Flanagan, "Multiplex vs. Multiple Selves: Distinguishing Dissociative Disorders," *The Monist* 82 (1999): 645–57.

needs, so others may have developed more than one self, in response to unusual circumstances, such as sexual abuse. What could be easier for a child already engaged in populating an imaginary world? And just as our imaginary playmates vied for the status of being our "best friend," so our imaginary selves may vie for the status of being our "true self." If so, then we suffer from MPD. Different selves take control at different times, but only in the same way as different imaginary friends succeed one another as favorite.

At this point, however, there is a gap in Dennett and Humphrey's account. When one imaginary friend supplants another as favorite, nothing much changes in the real world. But when one self supplants another in a patient diagnosed with MPD, the patient's behavior changes dramatically: he walks a different walk, talks a different talk, and expresses different states of mind. Surely, something has changed in the processes controlling his behavior.

Here is how Dennett and Humphrey explain changes of personality:

> The language-producing systems of the brain have to get their instructions from somewhere, and the very demands of pragmatics and grammar would conspire to confer something like Head of Mind authority on whatever subsystem currently controls their input. [...] Suppose, at different times, different subsystems within the brain produce "clusters" of speech that simply cannot easily be interpreted as the output of a single self. Then – as a Bible scholar may discover when working on the authorship of what is putatively a single-authored text – it may turn out that the cluster makes *best sense* when attributed to different selves. [SO, 42–43]

According to this explanation, different modules in the brain take control of the language-producing systems, yielding output whose interpretation calls for postulation of different Heads of Mind. Different selves thus correspond to different actual centers of control, but the selves are still fictional personifications of those centers, different *abstracta* postulated for the sake of interpreting a narrative containing severe discontinuities.

The problem with this explanation is that it accounts only for changes in the patient's verbal behavior, whereas multiples are reported to change their posture, gait, handwriting, and their projects and pursuits as well. Why should discontinuities in the patient's autobiography be accompanied by corresponding changes in the patient's course and manner of action? If a human being just contains "lots of subsystems doing their own thing," then why can't one of them do its thing with his feet even as

another does its thing with his mouth, so that he walks the walk of one personality while telling the story of the other?

An answer to this question is implicit in some of Dennett's descriptions of self-narration, but it attributes more sophistication to the self-inventor than Dennett acknowledges. The answer is that an autobiography and the behavior that it narrates are mutually determining.

In the case of the self-narrating robot, Dennett imagines a strict order of determination in one direction. He observes that "[t]he adventures of Gilbert, the fictional character, [. . .] bear a striking and presumably non-coincidental relationship to the adventures of this robot rolling around in the world."[19] And he explains this relationship between story and life by suggesting that the one is determined by the other: "If you hit the robot with a baseball bat, very shortly thereafter the story of Gilbert includes being hit by a baseball bat by somebody who looks like you." Presumably, the robot is designed to tell a story that corresponds to the life of that very robot.

What Dennett doesn't seem to imagine, in the case of this robot, is that he might also be designed to make his life correspond to his story. As Dennett tells it, the robot gets locked in a closet, calls out "Help me," and later sends us a thank-you note for letting him out. But surely a robot smart enough to thank us for letting him out of the closet would also be smart enough to tell us before he went back in. "I'm going into the closet" he would say, "Don't lock the door." And then he'd go into the closet, just as he had said he would. (If he didn't do what he had said, he might get stuck somewhere else and have to wait for help while we went looking for him in the closet.) A robot that can maintain correspondence in one direction, by saying that he's locked in the closet when he is, should be able to maintain correspondence in the other direction, by going into the closet when he has said that he will. Thus, whereas the robot will sometimes update his story to reflect recent events in his career, at other times he will narrate ahead of himself and then follow a career that reflects his story.

Although Dennett doesn't attribute this sort of sophistication to the robot, he does implicitly attribute it to a patient with MPD:

> Consider the putatively true case histories recorded in *The Three Faces of Eve* (Thigpen & Cleckley, 1957) and *Sybil* (Schreiber, 1973). Eve's three faces were

[19] CNG, 108. Note, then, that Dennett does not conceive of autobiographies as "entirely confabulated" narratives in which "anything goes" (Hardcastle and Flanagan, 650, 653).

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 the psychotherapists with whom I have talked about the case: When Sybil went in to see her therapist for the first time, she was not 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 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 the world with these new selves, more or less created on demand, under the eager suggestion of a therapist. [CNG, 111]

What does Dennett mean when he says that Sybil "engaged the world with these new selves"? Surely, he means that Sybil *acted out* the stories that she and her therapist had composed. She was a "living novel" in the sense that she not only narrated the roles she played but also played the roles that she narrated.

That's why Sybil's behavior always manifested the personality whose story she was telling at the moment. Her life shaped her story, and her story shaped her life, all because she was designed to maintain correspondence between the two. Hence the control of her speech and the control of her movements were not entirely independent. They were in fact *inter*dependent, since the controller of her speech must have been responsive to her movements, and the controller of her movements must have been responsive to her speech.

Yet if a self-narrator works in both directions, then the self he invents is not just an idle fiction, a useful abstraction for interpreting his behavior. It – or, more precisely, his representation of it – is a determinant of the very behavior that it's useful for interpreting.[20] Indeed, the reason why the narrator's representation of a centrally controlling self is so useful for interpreting his behavior is that it, the representation, really does control his behavior to some extent.

Of course, the central controller he has may not be much like the one he represents himself as having. After all, a self-narrator doesn't represent himself as being centrally controlled by his own story.

[20] Flanagan says, "[T]he self as represented has motivational bearing and behavioral effects. Often this motivational bearing is congruent with motivational tendencies that the entire system already has. In such cases, placing one's conception of the self into the motivational circuits enables certain gains in ongoing conscious control and in the fine-tuning of action" ("Multiple Identity, Character Transformation, and Self-Reclamation," in G. Graham and Lynn Stephens, eds., *Philosophical Psychopathology* [Cambridge, MA: MIT Press, 1994], p. 140).

Or does he?

In order to answer this question, we must consider some prior questions that Dennett overlooks. First, consider whether the behaviors attributed to Gilbert by the robot's novel-writing computer include the behavior of writing the novel. When the robot gets locked in a closet, he tells about Gilbert's being locked in a closet; but when he tells the story of Gilbert, does he also tell about Gilbert's telling that story? He says "Call me Gilbert"; but does he ever say, "I'm Gilbert and this is my story"? He writes a note that says "Thank you," but can he also write a note that says "I'm writing to say thanks"? I can't imagine why not.

Nor can I imagine how the robot would tell the story of Gilbert without including information about the causes and effects of the events therein. When he calls for help, he might well elaborate, "I've gotten myself locked in the closet," thus attributing his current predicament to what he did a moment ago. And when he writes his thank-you note, he might well begin, "I'm writing because you let me out of the closet," thereby attributing his present behavior to an earlier cause. A story that merely described one event after another, without mentioning any causal connections, would hardly qualify as a narrative.

Thus, the features of himself that the robot can ascribe to Gilbert ought to include this very activity of self-description; and he should also be able to describe the causes and effects of his activities, including this one. Hence in ascribing his activities to Gilbert, the robot should be able to describe the causes and effects of his doing so.

Now, what causal role might the robot attribute to his own remark, "I'm going into the closet"? He might say, "I'm telling you this because I'm on my way into the closet," thereby casting his speech as an effect of his movements. But this remark would be strictly accurate only if the robot was going into the closet anyway and was merely reporting on his current trajectory. What I have imagined, however, is that the robot goes into the closet partly because of having said so, in order to maintain correspondence between his story and his life. Insofar as the robot can report on the causes and effects of his behavior, then, he ought to say, "I'm going into the closet partly because I've just said so" – or, perhaps, "I'm hereby heading for the closet," a remark that implicitly ascribes this causal role to itself.

I think that human self-narrators make such remarks frequently, whenever they make promises or other verbal commitments, which may be as trivial as "I'm heading for the closet." As you putter around the office at the end of the day, you finally say, "I'm going home," not because you were

already about to leave, but because saying so will prompt you to leave. As your hand hovers indecisively over the candy dish, you say, "No, I won't," not because you weren't about to take a candy, but because saying so may stop you from taking one.[21] These utterances are issued *as* commitments, in the understanding that they will feed back into your behavior. Hence you do understand that your running autobiography not only reflects but is also reflected in what you do.

These observations suggest that the "central controller" of a person may indeed be a fiction, not in the sense that it is a fictional character in the person's autobiography, but in the sense that it *is* the person's autobiography – the reflective representation that feeds back into the person's behavior.[22] This central controller is in fact what social psychologists call the self. In the social-psychology literature, the word "self" denotes a person's self-conception rather than the entity, real or imagined, that this conception represents. And the same literature reports evidence for the feedback loop I have posited.

Researchers have found, for example, that subjects tend to predict that they will vote in the next election at a far higher rate than the average turnout; but that the turnout among those who have predicted that they will vote is also higher than the average.[23] Many who wouldn't otherwise have voted, it seems, end up voting because of having predicted that

[21] I discuss cases like these in "How to Share an Intention," *Philosophy and Phenomenological Research* 57 (1997): 29–50; reprinted in *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000).

[22] Dennett almost strays into this second conception of the self. For example:

> A self, according to my theory, is not any old mathematical point, but an abstraction defined by the myriads of attributions and interpretations (including self-attributions and self-interpretations) that have composed the biography of the living body whose Center of Narrative Gravity it is. As such, it plays a singularly important role in the ongoing cognitive economy of that living body, because, of all the things in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself. [RS 426–27]

> Dennett begins this passage by speaking of the self as an abstract object posited by the host's autobiography. But then he speaks of the self as playing "a singularly important role in the ongoing cognitive economy" of the host, and finally he describes it as "the model that the agent has of itself." At this point, it is unclear whether he is speaking of an abstract object or of the host's representation of it, which is a real element in the host's psychology, positioned to play a causal role in his mental economy.

[23] Greenwald, A.G., Carnot, C.G., Beach, R., and Young, B., "Increasing Voting Behavior by Asking People if They Expect to Vote," *Journal of Applied Psychology* 72 (1987): 315–18.

they would, thus conforming their lives to their stories.[24] Like Sybil, who "lived out" the novels that she composed with her therapist, these subjects lived out the predictions that they were prompted to make by the experimenters.

Similar research has documented a slightly different phenomenon, known as the attribution effect. Subjects can be led to act annoyed or euphoric depending on whether they are led to believe, of artificially induced feelings of arousal, that they are symptoms of annoyance or euphoria.[25] Subjects can be prevented from acting shyly in unfamiliar company by being led to attribute their feelings of anxiety to something other than shyness.[26] And researchers can modify the degree of retaliation that a subject carries out against putative aggressors by modifying the degree of anger that he believes himself to be feeling toward them.[27] All of these experiments suggest that people tend to manifest not just what they're feeling but also what they represent themselves as feeling. Whether they behave angrily depends, not just on whether they are angry, but on whether they interpret their feelings by updating their autobiographies with the attribution "I'm angry." Whether they behave shyly depends on whether the current episode of their autobiography says "I'm feeling shy."

Here the subjects are "living out" their self-conceptions in a more holistic sense. Unlike the self-predicting voters, they aren't doing things that they have described themselves as doing. Rather, they are doing things that would accord with what they have described themselves as feeling. But this process, too, is implicit in Dennett's account of self-narration.

[24] I explore this literature in "From Self Psychology to Moral Philosophy" (Chapter 10 in the present volume). For a more recent philosophical discussion of this phenomenon, see Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press, 2001), pp. 38 ff.

[25] Schachter, S., and Singer, J.E., "Cognitive, Social and Physiological Determinants of Emotional State," *Psychological Review* 69 (1962): 379–99.

[26] Brodt, S.E., and Zimbardo, P., "Modifying Shyness-Related Social Behavior Through Symptom Misattribution," *Journal of Personality and Social Psychology* 41 (1981): 437–49.

[27] Berkowitz, L., and Turner, C., "Perceived Anger Level, Instigating Agent, and Aggression," in *Cognitive Alteration of Feeling States,* eds. H. London and R.E. Nisbett (Chicago: Aldine, 1972), 174–89; Zillman, E., Johnson, R.C., and Day, K.D., "Attribution of apparent arousal and proficiency of recovery for sympathetic activation affecting excitation transfer to aggressive behavior," *Journal of Experimental Social Psychology* 10 (1974): 503–15; Zillman, D., "Attribution and Misattribution of Excitatory Reactions," *New Directions in Attribution Research,* vol. 2, eds. John H. Harvey, William Ickes, and Robert F. Kidd (Hillsdale, NJ: Erlbaum, 1978), 335–68.

For as we have seen, Dennett says that "[w]e try to make all of our material cohere into a single good story."[28] And acting in accordance with our self-ascribed emotions is a way of ensuring that our story-material will cohere.

Consider how this process might be implemented in the robot who calls himself Gilbert. If the robot is locked in the closet, his internal state may include the initiation of a subroutine that searches for avenues of escape from danger and quickly selects the one most readily available. This subroutine will have a name – say, "fear" – and so the robot will report "I'm locked in the closet and I'm starting to get frightened." And now two different modules in the robot will dispose him to take action. One is the fear module, which may recommend breaking down the door as one of several preferred alternative avenues of escape; the other is the narrative module, which will recommend "I'm breaking down the door" as one of several preferred continuations of the story. If after he said "I'm getting frightened," the robot continued his story with "I think I'll back up my hard disk," then he would no longer be writing a passable novel, since his "material" wouldn't cohere. His narrative module will therefore favor "I'm breaking down the door" as a more coherent way to continue the story. And the narrative module can go ahead with this continuation of the story, confident of being borne out by the robot's behavior, since the robot is sure to break down the door once his preexisting fear is reinforced, in motivating that behavior, by his disposition to maintain correspondence between his story and his life.

Thus, having attributed an internal state to himself ("I'm getting frightened"), the robot is influenced to act in accordance with that attribution. Like a human being, he tends to manifest fear not only because he's "feeling" it but also because he "thinks" it's what he's feeling.

I have now introduced the idea of the robot's having a "narrative module" that produces Gilbert's autobiography. This module must incorporate, first, the function of ensuring that the robot's story corresponds to its life and, second, the function of maintaining the internal coherence of the story itself. The module must be designed to produce a text that is both consonant with the facts and sufficiently consonant with itself to qualify as a story.

[28] CNG, 114, quoted on p. 205.

Moreover, I have suggested that the robot can maintain correspondence between its story and its life in either direction, by narrating its actions or by acting out its narrative. Hence in pursuit of narrative coherence, the module can sometimes choose, among possible turns in its story, the one that would best fit the story thus far, precisely because it can then influence the robot's life to take the corresponding turn. The narrative module needn't always depend on the robot's career to provide material for a coherent story; it can sometimes tell a coherent story and induce the robot's career to follow.

In previous work, I have argued that a creature equipped with such a module would amount to an autonomous agent.[29] I won't repeat those arguments here, but let me briefly illustrate some of them with the help of Dennett's self-narrating robot.

As Gilbert rolls down the hall, he may autobiographically announce where he is going. But he needn't just report where he is already programmed to go, since his disposition to maintain correspondence between story and life will dispose him to go wherever he says he's going. Suppose that he is in the middle of his Fetch New Batteries subroutine, which sends him to the supply closet (where he sometimes gets locked in). The fact remains that if he said "I'm on my way to the library," his disposition to maintain correspondence would dispose him to head for the library instead. So if another, concurrently running subroutine can get Gilbert's speech-producing module to emit "I'm on my way to the library," then it may be able to bring about a change of course.

Now, Gilbert's disposition to maintain correspondence wouldn't be sufficient to make him head for the library if no other subroutines inclined him in that direction. Even if he said "I'm on my way to the library," his Fetch New Batteries routine would still favor heading for the supply closet, and his disposition to bear out his story would be unlikely to override a routine for obtaining essential resources. But I imagine his inner workings to be in the following, rather complicated state. Various task-specific subroutines are running concurrently, and some of them are making bids for control of his locomotive unit, to propel him toward one destination or another. His Fetch New Batteries subroutine is bidding for a trip to the supply closet, while his Departmental Service subroutine may be bidding for a trip to the library, in order to fill a faculty member's

[29] See *Practical Reflection* (Princeton: Princeton University Press, 1989); and *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000).

request for a book. Meanwhile, the narrative-composing module is busy updating the story of Gilbert's most recent adventures and the ongoing evolution of his inner states, including which task-specific subroutines are running and where they are bidding him to go. And the disposition of this module to maintain correspondence between his story and his life, though not sufficient by itself to override other demands for locomotion, is sufficient to tip the balance in favor of one or another of those demands. So if Gilbert says "I'm heading for the supply closet," his disposition to bear out his story will reinforce the battery-fetching demands, and he'll head for the supply closet; whereas if he says "I'm heading for the library," his disposition to bear out his story will reinforce the demands of departmental service, and he'll head for the library instead. As long as the competition among those subroutines is not too lopsided, the narrative module is in a position to decide where Gilbert goes.

When I say that the narrative module can "decide" where Gilbert goes, I mean it can literally *decide.* For as we have seen, this module is in a position to have Gilbert speak the truth in naming any one of several destinations, each of which he would thereby head for, if he said so. The novelist in Gilbert can therefore *make up* where Gilbert is headed, choosing among different available turns in his story, none of which is privileged as the turn that the story must take in order to be true. As a self-narrator, then, Gilbert faces an epistemically open future – which gives him, in my view, as much free will as a human being.[30]

On what basis will the narrative-composing module make its decision? It can declare a winner in the contest among demands for locomotion, but on what basis will it adjudicate among those demands? The answer, already implicit in Dennett's theory, is that it will adjudicate on the basis of how best to continue the story – how to "make [its] material cohere."[31]

In many cases, acting on one demand will already make more narrative sense than acting on another, and the narrative-composing module will therefore declare a winner simply by telling the more coherent continuation of the story. But if neither continuation would make more narrative sense at this point, then the module can fill in more detail

[30] For a detailed defense of this claim, see my "Epistemic Freedom," *Pacific Philosophical Quarterly* 70 (1989): 73–97; reprinted in *The Possibility of Practical Reason.*

[31] CNG 114.

about its current situation, by recording which demand is stronger than the other or by recording more of the circumstances – which may arouse more internal states, which can in turn be recorded. At some point, the story will *become* more amenable to one continuation or other, and the narrative module can go ahead with the better continuation, thereby making its decision.

In this way, I believe, the module will decide on the basis of considerations that serve as reasons for acting. In canvassing Gilbert's outer circumstances and inner states, it will weigh them as considerations in light of which various possible actions would make sense. It will thus weigh Gilbert's circumstances and states as providing a potential *rationale* for his next action – that is, an account that would make the action intelligible, a coherent development in his story. When the novelist in Gilbert writes in the action with the best rationale, he will in effect be deciding for reasons.

Note that this claim places significant constraints on the conception of narrative coherence on which I can rely. One might have thought that whether an action would make for a coherent continuation of Gilbert's story ultimately depends on whether he has reason for taking it. My claim, however, is that whether Gilbert has reason for taking an action ultimately depends on whether it would make for a coherent continuation of his story. Because I make the latter claim, I cannot adopt the former in order to explicate narrative coherence, since my account would then become viciously circular: narrative coherence cannot ultimately depend on rational justification if rational justification ultimately depends on narrative coherence.

Of course, *we* can tell a story about Gilbert that makes sense because it portrays him as taking actions for which he has reasons; for we can portray him as taking actions because they cohere with *his* story. Indeed, I have already claimed that self-narration takes account of its own effect on the subject's behavior, by portraying him as *hereby* heading for the supply closet or the library. To this extent, self-narration already relies for some of its coherence on the fact that the subject is doing what coheres with this very story – hence on the fact that he is doing something for which he has reasons, as I conceive them. But this fact cannot be the sole basis for the narrative coherence involved. There must be some prior basis on which the subject's action makes sense in light of his story before it can also make sense in light of his tendency to do what makes sense.

The nature of narrative coherence is a topic that lies beyond the scope of this chapter.[32] But I have already indicated one basis on which Gilbert can regard actions as cohering with his story independently of his having reasons for taking them. I have supposed that Gilbert understands his own inner workings, in the form of the various subroutines that are vying to control his behavior. Gilbert understands that whatever he does will be controlled by one of these subroutines and will consequently make sense by virtue of having a causal explanation, which cites the relevant subroutine as the controlling cause. In considering which action would make for a coherent continuation of his story, Gilbert can look for an action that would have the most satisfying causal explanation in light of the subroutines vying for control.

Of course, where Gilbert has subroutines vying for control, human beings have conflicting motives, which serve as controlling causes of their behavior. Where Gilbert looks for an action that would best be explained by his subroutines, humans look for an action that would best be explained by their motives. That's why humans look to their motives – that is, to their desires and beliefs – as reasons for acting.

In deciding for reasons, the inner novelist plays the role that is ordinarily attributed to the self. A third conception of the self has therefore emerged. According to Dennett's conception of the self, with which I began, the self is the merely fictional protagonist of a self-narrator's autobiography. According to the second conception, the self is the autobiographer's reflective representation, which guides his actions as well as his speech. What has now emerged, however, is that control rests with the narrative module – the inner novelist, recording the subject's last step and declaring his next step, in a way that amounts to deciding for reasons. According to the third conception, then, the self is the narrator.

This third conception of the self no longer supports the skepticism of Dennett's initial conception. The protagonist of Gilbert's autobiography is no longer, as Dennett believes, a merely fictional character whose shoes cannot be filled by the actual author. Now that the robot has a central controller that makes decisions for reasons, he has a self, and so his story has come true.

Note that what fills the shoes of the protagonist in the story of Gilbert is the robot, not the robot's self. "Gilbert" is not the name of a self; it's the

[32] But see my "Narrative Explanation," *The Philosophical Review* 112 (2003): 1–25.

name of a unified agent who *has* a self, in the form of an inner locus of agential control. My current claim is that the self-narrating robot really is endowed with a self in this sense and can therefore live up to the portrait of the protagonist in his autobiography. He is endowed with a self because his inner narrator is a locus of control that unifies him as an agent by making decisions on the basis of reasons.

The self-narrating agent is a bit like an improvisational actor, enacting a role that he invents as he goes. The difference is that an improvisational actor usually invents and enacts a role that he is not playing in fact. His actions represent what they are not – actions other than themselves, performed out of motives other than his. By contrast, the self-narrator is an ingenuous improviser, inventing a role that expresses his actual motives in response to real events. He can improvise his actual role in these events because his motives take shape and produce behavior under the influence of his self-descriptions, which are therefore underdetermined by antecedent facts, so that he partly invents what he enacts.

Yet how can an agent act out invented self-descriptions without somehow falsifying them, by being or doing something other than is therein described? How can enacting a role fail to involve fakery or bad faith?

The answer is that when the agent invents descriptions to be enacted, he describes himself as the inventor-enactor of those descriptions. He describes himself as *hereby* heading for the supply closet or the library, thus describing his actions as flowing from these descriptions, as realizations thereof. The protagonist in his autobiography is therefore both fictive and factual – fictive, because his role is invented by the one who enacts it; factual, because it is the role of one inventing and enacting that role.

To be sure, a self-narrator can go beyond what is factual, if he applies self-descriptions whose autobiographical application won't make them true. Although he can sometimes tip the balance of his antecedent motives in favor of leaving the office by saying "I'm leaving," at other times he can't, and then a declaration of departure would be ineffectual – an instance of weakness of will. Alternatively, his motives for going home may already be sufficient to make him go home no matter what he says – in which case, "I'm leaving" is the only true thing for him to say. Within these constraints, however, the self-narrator retains considerable latitude for invention. Even if he is already determined to leave the office, he is probably capable of going home or going out for a drink, or perhaps just taking a walk, depending on what he writes into his story.

To this extent, I can endorse Dennett's claim that the self is a fictive character. Where I disagree with Dennett is over the claim that being fictive, this character doesn't exist in fact. Dennett thinks the real-life author of an autobiography is significantly different from the character portrayed as the protagonist. I think that the author of an autobiography is just like the protagonist, since the protagonist is portrayed as a self-improvising character, the inventor-enactor of his own story – or, as I prefer to say, an autonomous agent.

My disagreement with Dennett over the truth-value of a human being's autobiography results from two subsidiary disagreements. On the one hand, Dennett believes that a human being has no central controller, whereas I believe that Dennett himself is committed to crediting a human being with a central controller, in the form of a narrative intelligence. On the other hand, Dennett believes that a human being's autobiography portrays his central controller as a "brain pearl" or Cartesian ego, whereas I believe that this autobiography portrays the central controller as the narrative intelligence that it is. We live up to our aspirations with respect to selfhood, then, partly because we have more of a self than Dennett expressly allows, and partly because we aspire to less than he thinks.

I have overlooked another disagreement with Dennett, which I should mention before closing. Although Dennett tries to deny the unity of the self-narrating agent, he commits himself expressly to the unity of the narrative – to the proposition that "We try to make all of our material cohere into a *single* good story."[33] Indeed, the unity of this narrative seems to account for the temporal unity of the purely fictional self in which Dennett believes. This fictional character remains one and the same self because he is the protagonist in one and the same continuing story.[34]

In my view, however, we tell many small, disconnected stories about ourselves – short episodes that do not get incorporated into our life-stories. The process of self-narration shapes our day-to-day lives in units as small as the eating of a meal, the answering of a phone, or even the scratching of an itch; but our life stories do not record every meal eaten,

[33] CNG 114 (quoted on p. 205), emphasis added.

[34] This view is endorsed by Flanagan, "Multiple Identity," p. 136: "Augustine's *Confessions* is an autobiography. It is the story of a single self. This is established in part because Augustine is able to produce an account that narratively links up the multifarious episodes of his life from the first-person point of view."

every phone answered, or every itch scratched. Because the narratives of these minor episodes are never unified into a single story, their protagonist cannot derive his unity from theirs. The agent who types this letter 'a' is the same person who cut his forefinger with that pocketknife in the summer of 1959, but not because there is any single narrative in which he figures as the protagonist of both episodes.

So when I describe the inner narrator as a unified self, I am not speaking of the temporal unity that joins a person to his past and future selves; I am speaking of agential unity, in virtue of which a person is self-governed, or autonomous. In my view, autonomy is not related to personal identity in such a way that a single entity plays the role of self in both phenomena: that which makes us self-governed is not that which makes us self-same through time.[35]

[35] I argue for this view in "Identification and Identity" (Chapter 14 in the present volume).