Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK, and 350 Main Street, Malden, MA 02148, USA
METAPHILOSOPHY
Vol. 36, No. 4, July 2005
0026-1068

NATURAL FREEDOM

DANIEL C. DENNETT

Abstract: Three critics of *Freedom Evolves* (Dennett 2003) bring out important differences in philosophical outlook and method. Mele's thought experiments are supposed to expose the importance, for autonomy, of personal history, but they depend on the dubious invocation of mere logical or conceptual possibility. Fischer defends the Basic Argument for incompatibilism, while Taylor and I choose to sidestep it instead of disposing of it. Where does the burden of proof lie? O'Connor's candid expression of allegiance to traditional ideas that I reject highlights a fundamental difference in assumptions about how—and why—to do philosophy. There are indeed definable varieties of free will that are incompatible with determinism. Do they matter? I have argued, against philosophical tradition, that they don't.

Keywords: autonomy, causation, determinism, freedom, naturalism.

Dearly beloved, I want to thank Brother Tim O'Connor for his candid reactions to my published sermons, and I welcome you all, in the spirit of ecumenicism, to the Church of Fundamentalist Naturalism. Let me tell you a bit more about the church. Our symbol is of course the Darwin-fish, the four-legged evolver that echoes the ancient fish symbol of Christianity. I was wearing my Darwin-fish lapel pin at an evolutionary theory conference a few years ago, and the physicist Murray Gell-Mann came up to me and after reminding me of what he regarded as the first known acronym—IXΘYΣ, the Greek word for *fish*—Ιησους Χριστος Θεου Υιος Σωτηρ, Jesus Christ, God the son and savior—he asked me what D-A-R-W-I-N stood for. I said I'd get back to him and went off to have a cup of coffee while dredging up what I could of my high-school Latin. I came up with something I'm quite happy with: *Delere Auctorem Rerum Ut Universum Infinitum Noscere*: Destroy the Author of things in order to understand the infinite universe! That, it seems to me, is our key insight about Darwin's dangerous idea: in a single nonmiraculous stroke, it unites the world of meaning and purpose and design with the world of meaningless matter and mechanism, cause and effect.

Let the service begin with the Rite of Confession. Brother Al Mele gently notes in his essay that I have committed an unusual sin; *anti-*

plagiarism: I misattribute to him and his book *Autonomous Agents* (1995) an invention of someone else, the Principle of Default Responsibility. Dumfounded by his claim, I returned to my file of notes on his book, made when I read it some years ago, and sure enough, there is the Principle of Default Responsibility, capital letters and all, but apparently part of my musings on his book, not a quotation. So Al is the muse, not the author, of the principle, and I apologize for my error. *Mea culpa*. Had I realized that it was my own formulation when I put it into my book, I might have tinkered with it, expanded or enlarged upon it, but there it is, for others to make of what they will. I made it up; no wonder I approved of it. I'm just glad I didn't commit the much more serious sin of foisting a principle onto him that I then subjected to withering rebuttal and scorn! He also notes another instance of sloppy scholarship on my part. I overlooked, as he says, the fact that in his tale of Ann and Beth he had specifically headed off my objection about pseudomemories: How did I overlook this? Well, this is what he says about Beth's brainwashing: "Beth did not consent to the process. Nor was she even aware of it" (Mele 1995, 145). Doesn't her lack of awareness of the brainwashing entail some pseudomemories in its place? Perhaps not. But Al is right that he explicitly says Beth is surprised by the changes in her values, so apparently Beth isn't as radically misinformed as I imagined her to be. Once again, *mea culpa*. Let's see what difference this makes.

As Al notes, I might be expected to find his historical condition appealing, given some of the things I say about how we make ourselves into responsible agents over time. Why don't I go along with his historicized version of compatibilism? My view about the role of history here is the same as my view about history in other philosophical contexts: it is indeed important, because *in the actual world* only certain sorts of historical processes can create what *really* matters: the set of competences implicit in the total structure of an embodied nervous system. The sorts of philosophical thought experiments that try to tease apart the history from the current state all depend on "logically possible" but preposterous scenarios. My parody is the case of the cow-shark.

> Suppose a cow gave birth to something that was atom-for-atom indiscernible from a shark. Would it be a shark? If you posed that question to a biologist, the charitable reaction would be that you were making a labored attempt at a joke. . . . Is a cow-shark a shark? It swims like a shark, and mates successfully with other sharks. Oh, but didn't I tell you? It is atom-for-atom indiscernible from a shark, except that it has cow DNA in all its cells. Impossible? Not *logically* impossible (say the philosophers). Just so obviously impossible as to render further discussion unnecessary. (Dennett 1994, 518; 1996, 76–77)

The philosophical problem of cow-sharks is not worth discussing. Swampman is less obviously bogus, but still bogus, for the same reasons —or at least so I argue. But aren't Al's examples of brainwashing much

more realistic, much closer to the realm of genuine possibility? That is the issue. Let's consider the new case, of Ted and Fred, that Al mentions in his article for this symposium:

> Fred truly believes that he has not been brainwashed whereas Ted falsely believes this about himself. Imagine that after Ted has lived with and thought about his new values for a few hours, and before he takes out the loan, we show him a video of his being brainwashed and persuade him of the truth about how he acquired his new values. He is now no less informed than Fred, and *I do not see why* [my italics] his now having this true belief must render him less flexible, less open-minded, or less capable of resisting temptation than he was just before he acquired it.

Strangely enough, what Al himself cannot see, his imaginary interlocutor Pad surmises with ease: "You have a nonchalant attitude toward such a drastic change in beliefs. That change would definitely be psychologically devastating to Fred and to any possible being who had been acting freely, and the devastation would definitely render any such being unfree—at least for quite some time." Of course Pad is talking of a *slightly* different case, in which the autonomous Fred is given the misinformation about being brainwashed, but Al's response to his own interlocutor has an ominous ring to it: "Now is it *conceptually impossible* [my italics] that Fred absorbs the shock well enough to take out that loan for his daughter?" No, I daresay it isn't conceptually impossible, in the philosophers' strained sense of the term. But so what? It also isn't *conceptually* impossible that Beth should find herself with unshedable values in spite of their unanticipated and inexplicable overnight arrival, while still preserving her basic sanity. But it is, in fact, impossible.

Al and I agree that history is important; the only thing we disagree on is whether it is important *in itself* aside from the effects it produces so reliably. He even agrees with me, I think, that the reason history is important is that the right sorts of histories yield just the sort of informedness, open-mindedness, and flexibility that I make the hallmark of autonomy. So the only thing that divides us, so far as I can see, is how to handle cases that could never come up: Swampman-type agents who are only apparently autonomous according to him but fully autonomous according to me.

He notes that "all bets are off" if we tweak the thought experiments further and allow Ted to have initiated the brainwashing process himself. I think this is right, and important. The implications of these prospects are elegantly surveyed by Peter Suber (2001). Finally, I must say that I myself am going to find it difficult in the future not to think of my book as *Elbow Pad*.

Let me now turn to John Fischer's essay. In the closing pages of *Freedom Evolves*, I offer a passage in which I lead with my chin:

> I have not just lavished attention on the ideas of non-philosophers; in the process I have ignored the ideas of more than a few highly regarded philosophers, sidestepping several vigorously debated controversies in my own discipline without so much as a mention. To the participants in those debates I owe an explanation. Where, some may well ask, are my refutations, my proofs, my philosophical arguments demonstrating the unsoundness of their carefully crafted analyses? I have provided a few: Austin's putt, Kane's faculty of practical reasoning, and Mele's autonomy, for instance, have come in for the sort of detailed attention philosophers expect. With regard to others, I have decided to put the burden of proof on them. It takes a certain amount of shared background assumptions to make a philosophical controversy, and I have convinced myself—not proved—that my informal tales and observations challenge some of their enabling assumptions, rendering their contests optional, however diverting to those embroiled in them. I could have said exactly how and why, but it would have taken a hundred pages or more of dense textual exegesis and argument, ending up with verdicts of false alarm, an anticlimax to be eschewed. That is a risky decision on my part, since it is open to them to demonstrate that I have woefully underestimated the inevitability of their shared presuppositions, but it is a risk I am prepared to take. (Dennett 2003, 307)

Quite appropriately, John rises to the occasion and claims that I have *not* managed (even with lots of help from my coauthor, Christopher Taylor) to slide by the thickets of work on what he calls the Basic Argument for incompatibilism. In the book, I don't even discuss the Basic Argument, but in the essay, Taylor and I provide an appendix in which we briefly discuss a version, essentially van Inwagen's, and say what we think is wrong with it. This is where John has something to get his hands on that is more meaty than my insouciant disregard for this literature. He sees several problems with our handling of this version of the Basic Argument, all hinging on how to understand the term *cause* in

> 3. If A has the power to cause α and $\alpha \Rightarrow \beta$ obtains in every possible world, then A has the power to cause β. (Taylor and Dennett 2001, 273)

This is the premise we deny, following our recommended policy of treating counterfactual necessity, not sufficiency, as the heart of the (appropriate) concept of causation. John agrees that if we stay with our concept of causation, premise 3 has no warrant, but he urges that "there is a weakly causal reading" that makes premise 3 defensible. He does not claim that this weakly causal reading "captures some ordinary, commonsense idea," saying rather that it "builds on such ideas but departs slightly from them to create a theoretical notion that can then be deployed in the Basic Argument." And—no surprise—when we substitute John's reading for our recommended reading, the Basic Argument is "not obviously problematic, and certainly not problematic in the way indicated by Taylor and Dennett."

So, just to review the bidding, Taylor and I argue for a necessity-based notion of causation, and we show (in passing, in an appendix) how it enables us to sidestep the notorious Basic Argument that has troubled philosophers for thousands of years—by John's reckoning. He agrees that *given our notion of causation*, the Basic Argument falls apart, but he proposes his own "theoretical" notion of causation, acknowledges that it does not have even the imprimatur of ordinary language but does have the virtue of saving the Basic Argument for another day, or maybe another millennium. We wouldn't want to say farewell to something as much fun as the Basic Argument in an appendix would we? Well, yes.

John also disapproves of our cavalier treatment of events and sentences. "This tends to conflate issues about causation of events and causation of states of affairs or propositions or the truth of sentences, that is, causing it to be the case that a certain state of affairs obtains (or a certain proposition or sentence is true)." These issues "are complex and require delicate analysis. I do not believe that the Taylor and Dennett account is sufficiently nuanced to handle such complex states of affairs, and their conflation . . . leads them astray here." Perhaps he is right, but then now would seem to be the ideal time for him to sock it to me. Just how does our simpler view lead us astray? Please show us all a serious gap in our account, a distinction we will want to honor and can't, a case we can neither handle nor dismiss as ignorable. That is how one would go about countering the suspicion that all this complexity and delicacy that Taylor and I walk past may be an artifact of the controversy, technological innovations spun in a philosophers' arms race but having no peacetime use. Perhaps John is merely being polite in refraining from going into the details of our muddlement, but if the only virtue of all that complexity is that it permits the continuing discussion of the Basic Argument, I will continue to be unembarrassed by my refusal to grapple with the fine details of that literature. He says that Taylor and I "cannot dispose of the 'modal' or 'transfer' version of the Basic Argument so easily," and he is surely right: we could not *dispose of* all the versions of that argument without showing how each and every one of them is mistaken. And that is a daunting challenge, the work of a lifetime. But having given and defended our own reading of "cause" and shown that *it* does not lend itself to the argument, we *can* quite easily excuse ourselves from the onerous task of *argument-disposal* until such time as we are shown to have left out something important.

Showing exactly that is what John next attempts to do. He does so by setting aside the appeal to any such "transfer principle" as premise 3 and appealing instead to the "extremely plausible and intuitively attractive Principle of the Fixity of the Past and Laws: an agent has it within his power to do *A* only if his doing *A* can be an extension of the actual past, holding the natural laws fixed." He says that his argument using this Fixity Principle is "a version" of the Basic Argument, but since one of the

points of the Taylor and Dennett essay was to disentangle the two issues of causation and possibility, I have to view this claim as an instance of the conflation we were exposing.

John says that his (extremely plausible and intuitively attractive) Principle "captures Carl Ginet's point that our freedom is the freedom to add to the given past (holding the natural laws fixed)." And he adds that if one accepts this fundamental idea, then one can state an argument for incompatibilism that is invulnerable to Taylor and Dennett worries. I'm sure he is right, but I do not accept this fundamental idea. That is precisely why I was so pleased with Christopher Taylor's way of showing what is wrong with it: it insists on taking the "actual past" as "given," down to the last electron, but it simply neglects the fact that this is an *unmotivated insistence*. When we are interested in whether somebody could have done otherwise—the phrase John himself uses in the conclusion of his alternative argument—we are *never* interested in this way of construing the issue, unless we are doing philosophy and confronting the Basic Argument!

I don't deny that there is a definable concept of freedom that is incompatible with determinism; I just deny that anybody should care whether one is free in that sense. If you really want to "extend the given past" in the way defined by Ginet, then if determinism is true, you are in for a disappointment. But why should you care? This metaphysical condition has nothing to do with *talent*, let alone such highly charged issues as personal responsibility, authorship, and moral character.

John agrees with at least most of this. He says in his review of *Freedom Evolves* in the *Journal of Philosophy*: "Both of us hold that [libertarian freedom] is not necessary for moral responsibility. But I take seriously the libertarian's wish to have such freedom" (Fischer 2003, 634). I don't. That is, I am myself utterly unmoved by the libertarian's wish, and I wonder if it is due to some confusion or other—most likely the confusion Taylor and I point out. In the end, John agrees that "our pragmatic interests and epistemic situation force an interest in broad possibility [the Taylor and Dennett kind of possibility]" but goes on to add, "but that interest is completely compatible with the idea that our freedom is the freedom to add to a given past, holding fixed the laws of nature. It is also completely compatible with the idea that what is possible, in the sense that is relevant to our planning, occurs along future paths that branch off the present, holding the past fixed." I grant the mere compatibility; what I am still looking for is a reason why the libertarian's wish should galvanize us any more than the idle metaphysical desire that one have a doppelgänger outside one's light cone. I suppose it would be cool, but I'm not going to worry over arguments for or against its truth.

This brings me, finally, to Tim O'Connor's reactions to my naturalism, which I find as usefully challenging as they are amusing. "Chances are, . . ." he says at the outset, "you're not a member of Dan's church, either."

This is to me a most interesting point: Tim, for one, is sure that the APA is not teeming with naturalists, and if he is right, then I submit that philosophy, as a discipline, has a serious problem. Because my brand of naturalism, however quaint and "fundamentalist" it may seem to some philosophers, is pretty much the unargued ambient worldview of the scientific community, and in case you haven't noticed, the denizens of that community outnumber us philosophers by better than a hundred to one. Those are just numbers, of course, and *majority* doesn't *rule* in any serious investigation, but at least one would expect philosophers to take seriously the metaphysical and epistemological presuppositions—however unexamined—of such a highly informed and successful group of researchers. As Tim says, my self-appointed role is largely that of drawing out some of the more unsettling implications of this shared naturalism, and one can go one of two ways with the results: are they an unintended reductio ad absurdum of this naturalism, or a measure of the amount of counter-intuitive adjustment that we will *all* have to make in the coming years to the further developments of this view? (Philosophers of Tim's persuasion might take comfort in knowing that one of the reactions I often get from scientists to my arguments is that they thought they were good mechanist materialists until they contemplated what I told them they would have to give up to be consistent, and now the attractions of dualism look better than they would ever have imagined!)

One of Tim's best jokes in a generally funny article is this: "What makes a naturalist a fundamentalist? Naturalists are an educated bunch (as fundamentalists often enough are), but their intellectual diet is narrow." Right. Compared with the intellectual diet of people doing analytic metaphysics, for example! I don't think Tim believes this, but since he raises the issue, I cannot forbear commenting. There seems to be a trend among today's graduate students in philosophy to return to the blissful know-nothing attitude of my graduate-student days during the heyday of ordinary-language philosophy. "I was attracted to philosophy as a career," a promising graduate student candidly told me last year, "because I saw that I wouldn't have to learn anything much about anything!" Shades of Oxford in the mid-1960s, when all you needed was a well-thumbed copy of *Philosophical Investigations* and a talent for glib ripostes and counterexamples. I must say that I view anybody in our profession who encourages this attitude in students, when there are so many wonderful opportunities to do serious research on important topics, to be guilty of a serious crime of intellectual seduction. True, there have always been a fair number of people in our field who have had no project beyond exhibiting and exercising their own cleverness, but I for one do not think this is a tradition we should honor.

Back to Tim's interpretation of my "fundamentalist" strictures on philosophical method: "In thinking about human beings and our place within the wider scheme of nature, we are not entitled to make empirically

'risky' assumptions in advance of hard evidence that has the imprimatur of a mature science—however well certain assumptions may fit our prereflective view of ourselves." No, of course you're *entitled* to do it, and I suppose there's no harm in it, so long as you know what you're doing. In such an exercise, you work out the logical implications of the prereflective intuitions of your informant—who is yourself. Back in the 1970s, the artifical-intelligence researcher Patrick Hayes embarked on a project to formalize a portion of folk physics or what he called *naive* physics—the physics we all betray knowledge of in our everyday life: towels absorb water, shadows can be projected through clear glass, things drop when you let go of them and often bounce (depending on the surface they land on), when things collide they make a noise, and so forth (Hayes 1978). We literally couldn't live without naive physics; it is extremely swift and fecund in its deliverance of reliable expectations, and virtually involuntary. You can't readily turn off your expectations. For instance, you "unthinkingly" leap back from the table when a water glass is overturned, expecting the water to roll off the edge onto your lap. *Somehow* your brain generates that expectation from its current perceptual cues and takes the usually appropriate avoidance action. The background machinery of naive physics is not directly accessible to introspection but can be studied indirectly by mapping its "theorems," the generalizations it can be seen to endorse (in a manner of speaking) by its particular deliverances. Many magic tricks exploit our intuitions of naive physics, gulling us into overlooking "impossible" possibilities, or inducing us to jump to conclusions (unconsciously) on the basis of a perceptual cue of one sort or another. Then there are the counterintuitive phenomena that baffle us naive physicists: gyroscopes, pipettes (why on earth doesn't the soda fall out of the bottom of that straw—it's wide open!), siphons, sailing upwind, and more.

Hayes's delicious idea was to try to formalize a portion of naive physics, the naive physics of liquids, yielding an explicit *theory* that would predict all the things we actually expect from liquids and hence predict *against* the things liquids do that we view as anomalies, such as siphons. Siphons are "physically impossible" according to naive physics. (This was a classic exercise in GOFAI, John Haugeland's Good Old-Fashioned Artificial Intelligence [Haugeland 1985], or what I have called the paradigm of the Walking Encyclopedia [Dennett 1991]; if a robot is to mimic a normal person's "knowledge of folk physics" it will have to have a version of that theory written in its memory banks, with a swift-inference engine attached, so that expectations can be generated in time for action guidance.)

What Hayes set out to do was a kind of rigorous anthropology, attempting to axiomaticize the false theories found among the folk. Let's call it *aprioristic anthropology of naive physics*, to mark its resolute refusal to let the actual facts of physics get in the way of deducing the

implications of its found axioms. The physics was naive, but Hayes was not. His project was *sophisticated aprioristic anthropology*, since he was fully alert to the fact that false theories are just as amenable to formalization as true theories, and he withheld all allegiance to his axioms. One could attempt just the same sort of project with folk psychology: deducing the implications of whatever is deemed "axiomatic" (unquestioned, impossible to deny, too obvious for words) by the folk. Call this enterprise *sophisticated aprioristic anthropology of folk (naive) psychology*, and contrast it with *naive aprioristic anthropology of folk (naive) psychology*, whose practitioners would make the mistake of committing themselves to the beliefs of the folk. And if one's set of informants is the singleton *oneself*, you get *naive aprioristic autoanthropology of folk psychology*. (The preceding three paragraphs are drawn, with revisions, from Dennett 2005.)

Now my question is: How if at all does "traditional" philosophy of mind, of the sort now making something of a comeback in the anti-cognitive-science backlash, differ from this enterprise? I don't detect any differences in either methods or sources of data.

Tim announces that he is "committed to causal realism," the doctrine he explicates thus: "Some things make other things happen, and the truth makers of these causings are not to be gleaned from suitable regularities in patterns between isolated bits of inert facts." This is an interesting and somewhat compelling vision of causation, but why be *committed* to it? David Lewis, Jaegwon Kim, and their many ingenious students have explored the ins and outs of our everyday notions of causation, trying to find a consistent folk theory thereof, and I submit that the way to understand this enterprise, known in our discipline as analytic metaphysics, is as a kind of anthropology, educing the best possible version (most perspicuous, consistent if possible) of some current folk concepts. If it is like good anthropology, it reserves judgment about the ultimate soundness or utility of the folk theory, an assessment that requires folk theory to be put into registration with the best scientific theories or, failing that, put into competition with such theories—an inquiry few metaphysicians are well equipped to conduct. (For an example of such a scientifically sophisticated theory of causation, see Judea Pearl 2000.)

Tim goes on to criticize my discussion of Austin's notorious putt, and the lessons to be drawn from it: "Here I think Dan fails to appreciate that the all-in sense of *can* is not disjoint from the ability sense but is rather stronger than it. It is the ability sense *plus*. . . . No experiments that we are able to undertake could confirm the further condition of (robust) causal indeterminism—we simply assume that in practical life." But since the assumption has no testable consequences, we might well ask: Are we right to assume this? Need we assume this? Is it perhaps at best merely an enabling or life-enhancing myth that we may usefully profess but need not take seriously, like the situation in Garrison Keillor's imaginary town of

Lake Wobegon, "where all the children are above average"? If the all-in sense of *can* has no other role to play, and it alone is what is motivating incompatibilism, we do well to see if we can get along without it.

Later Tim asks: "Why think that the very same notions of avoidance/avoidability are at work when we assess impersonal systems, even sophisticated, information-processing ones, and human beings to whom we attribute freedom and responsibility? It seems rather that we have at least two important senses." Perhaps. But is the second sense really important? Why? Again, I am not claiming that there is no such other sense (though I haven't seen the case yet made that there is) but am simply demanding that before any such sense is detached from its bracketed home in autoanthropology and put to work, it needs to be *motivated* by something beyond tradition; its role needs to be defended on the further ground that it makes a real distinction that science must honor.

Tim goes on: "For a person to be responsible for what he or she does, his or her causal activity must not entirely consist in the activity of impersonal constituents." This is, as near as I can tell, a frank avowal of some kind of Cartesian dualism. While I find the anthropological claim that at least some schools of folk (moral) psychology are committed to dualism to be somewhat plausible and worthy of defense, and while, in my role of anthropologist or heterophenomenologist, I can hardly vouchsafe the *autoanthropological* claim from Tim that he himself is committed to dualism, as a move in the debate over compatibilism and incompatibilism it falls short. He advocates "ontologically emergent" capacities, exhorting: "Think new basic capacities resulting from irreducible properties of whole systems!" Go ahead, but while you're at it, you'd better try to give scientists a way of taking these properties seriously. "Dan doesn't seem to like it when a philosopher shrugs his shoulders and says, 'Those are all fascinating questions for empirical researchers. It would be great if neuroscience can attain a degree of understanding of . . . human brains. . . . '" He's right. I don't like it when philosophers pass the buck in this way, since what philosophers sometimes seem not to realize is that it isn't enough to contrive a definition of a putatively logically consistent capacity; you have to motivate it in more than mere conceptual terms. Otherwise you are just positing what I have called "wonder tissue." There may be wonder tissue, undreamt of to date by the relevant scientists, but we need a better reason to hunt for it than that your intuitive theory of responsibility demands it. One way of looking at what I am doing in *Freedom Evolves* is to execute a sort of squeeze play, showing that more and more and more of what your pretheoretical intuitions insist on can in fact be provided for without the wonder tissue that tradition imagines. Still, you say, you want wonder tissue. Why? Just for old times' sake?

A reader of my book, Pamela Robinson of Citrus Heights, California, recently sent me a gift—a glorious bit of needlework on which she had embroidered a little poem from her childhood:

Due to Circumstances
Beyond my Control,
I am the
Master of my Fate
and the Captain
of my Soul.

Amen. This nicely captures one of my main themes, and I plan to adopt it as one of the canonical texts in the Church of Fundamentalist Naturalism.

Center for Cognitive Studies
Tufts University
Medford, MA 02155
USA
daniel.dennett@tufts.edu

References

Dennett, Daniel C. 1991. "Mother Nature Versus the Walking Encyclopedia: A Western Drama." In *Philosophy and Connectionist Theory*, edited by W. Ramsey, S. Stich, and D. E. Rumelhart. Hillsdale, N.J.: Erlbaum.

———. 1994. "Get Real." *Philosophical Topics* 22:505–68.

———. 1996. "Cow-sharks, Magnets, and Swampman." *Mind and Language* 11:76–77.

———. 2003. *Freedom Evolves*. New York: Viking Penguin.

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

Fischer, John Martin. 2003. "Review of *Freedom Evolves*." *Journal of Philosophy* 100:632–37.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: MIT Press.

Hayes, Patrick. 1978. "The Naive Physics Manifesto." In *Expert Systems in the Microelectronic Age*, edited by D. Michie. Edinburgh: Edinburgh University Press.

Mele, Alfred. 1995. *Autonomous Agents*. Oxford: Oxford University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Suber, Peter. 2001. "Saving Machines From Themselves: The Ethics of Deep Self-Modification." Available at http://www.earlham.edu/~peters/writing/selfmod.htm.