Available online at www.sciencedirect.com

SCIENCE DIRECT®

Artificial Intelligence 169 (2005) 165–173

Artificial Intelligence

www.elsevier.com/locate/artint

Book review

D. Dennett, Freedom Evolves, Viking, 2003.

Free at last! Free at last! Thank evolution, free at last!

Drew McDermott

Department of Computer Science, Yale University, P.O. Box 208285, New Haven, CT 06520-8285, USA

Available online 3 November 2005

I like almost everything Dan Dennett writes, and this book is no exception. Its topic is free will and issues generally linked with free will, such as the foundations of ethics, all treated from an evolutionary perspective. The reason for this perspective is Dennett's belief that all explanations of the way humans are must be grounded in an explanation of how evolution made them that way. This belief has been especially obvious since the publication in the mid-90s of his book *Darwin's Dangerous Idea*, but one finds frequent references to evolution in earlier works of his as well.

Dennett wrote a previous book on free will, called *Elbow Room*, almost twenty years ago. His reasons for returning to the subject are, apparently, some new technical results on determinism and inevitability, and a desire to focus more explicitly on Darwinian explanations of human moral sentiments. The new book does not represent a change of opinion since the earlier one, however, and sometimes I found myself wondering whether a new book-length treatment was appropriate. There are several places in the book where footnotes acknowledge previous Dennett papers from which pages or paragraphs have been adapted. The style of the book is not as polished as we are used to. There is an imaginary interlocutor named "Conrad" who is supposed to bring up objections for Dennett to refute, but Conrad makes only sporadic appearances, so the device seems a bit wobbly. Terms are borrowed from previous books without sufficient explanation. In *Darwin's Dangerous Idea*, the terms "crane" and "skyhook" are introduced carefully. A *skyhook* is an explanatory mechanism that hangs loftily in the air, whose only defect is that nothing supports it but itself; which is a pity, because its exalted height allows it to lend luster to what it

E-mail address: drew.mcdermott@yale.edu (D. McDermott).

doi:10.1016/j.artint.2005.10.002

explains. A *crane* is a workaday mechanism that explains by showing how the entity to be explained (such as morality) can be arrived at by a steady series of erections from solid ground; its seeming defect is that by building something noble (such as morality) out of crass ingredients, it makes the noble crass. (Whether this seeming defect is really a problem depends on the details. In the cases Dennett focuses on, he argues that it never is.) In *Freedom Evolves*, the terms "crane" and "skyhook" appear abruptly, with one sentence of explanation, and then are used in a few crucial contexts where they will mystify the uninitiated.

These problems are relatively minor for the true-blue fan, who will be pleased with the usual smooth style and clever insights to be found in Dennett's work. In fact, the only problem I have with his books and articles is that when someone asks me what he said, I end up telling them *my* thoughts on the topic. This is not just megalomania on my part; if I go back and scan the text, I realize that all its brilliant pieces are hard to distill into a quick summary, so people that agree tend to summarize by saying what *they* think, and people who disagree tend to dismiss by describing a caricature.

So let me carefully review what Dennett is saying in this book and try hard not to put words in his mouth. There are ten chapters, which seem to me to be divided into two parts, although Dennett makes no such division. Part 1, Chapters 1 to 4, is devoted to the technical question of whether determinism is compatible with free will; or, more generally, whether the facts of physics have any bearing on whether we are free. Part 2, Chapters 5 to 10, is about what sort of moral responsibility really exists in a world where mind is implemented using physical mechanisms. The first part is more technical, that is, directed more toward the philosophical community. The second is more a pseudo-historical sketch of how moral responsibility might have evolved, and what might threaten its continued existence.

Chapter 3 is the most interesting technical chapter.[1] Its main target is the intuition that in a deterministic world all events have "causes" that make them "inevitable". This intuition does seem undeniable, and Dennett makes a fascinating case that it is false. The key idea is that when we ask for the cause of an event, what we are asking for is a prior fact or event that is a necessary and sufficient condition for the later event to have happened. This is way too simple, and there is a huge philosophical literature devoted to cases in which A shoots B, but B would have died anyway because C shot him a second after A did. However, the simple formulation is enough to show the fallacy involved in states of a deterministic world as causes of events. For one thing, although a prior state is sufficient for all later events, it is not necessary. Some later events would have happened even in alternative worlds where the universe was never in precisely that state. But a more interesting reason is that we simply don't accept a prior state as the cause of an event we care about. For example, the question of what causes wars is of great interest to everyone interested in peace. No one would accept "The state of the universe on January 1, 1900" as the answer to the question "What caused World War I?" The point is that in talking about causes of an event, we are interested in situations that are similar, not just those that are identical, to the predecessors of the event. The reason is that we are interested in what to do in future situations, which

[1] It is based on work with Christopher Taylor [5].

we know will be different from past ones in many ways, but may well be similar in ways relevant to our interests.

Chapter 4 is a detailed look at theories of free decision making that depend on a particular kind of atomic decision that is absolutely uncaused, existing, it is fashionable nowadays to claim, in quantum-mechanical gaps in the causal flow. Dennett does an excellent job of breaking this theory down into useless splinters. The closer you examine these crucial "self-forming" decisions, the less weight they seem to actually carry. In the end there is no morally relevant distinction between an agent that makes truly indeterminate decisions and one that is driven by a deterministic pseudo-random number generator.

Part 2 of the book changes course rather dramatically, and focuses on how moral responsibility fits in with evolutionary theory. It must start with the question of how conscious decision-making creatures evolved, which turns out to be a long story. In Chapter 5 the focus is on the evolution from inert molecules to "situation-action machines" to "choice machines" (quoted terms are due to Drescher [1]). A situation-action machine is one that follows a set of rules in situations its sensors reveal. A choice machine is one that weighs the overall consequences of different actions in a given situation and tries to choose the one that maximizes utility. Chapter 6 brings in a favorite theme of Dennett's, the *meme*, a pattern of thinking, usually expressed with words, that reproduces itself by being copied from brain to brain in a way analogous to the way genes are copied. Memes are central to Dennett's theory of evolution of mind because they allow language to be the "crane" that erects consciousness out of intelligence. Consciousness is a way an intelligent being models itself, and in Dennett's opinion the step to modeling ourselves that way is a matter of communication and culture, mediated by memes.

Chapter 7 raises the question of altruism in evolutionary theory. Classical evolutionary theory suggests that altruists will be outplayed in the game of life by cheaters who take the altruist suckers for all they've got. A nice of way of fixing this bug is to suppose that even though individuals always do better by cheating, *communities* might do better by harboring mostly altruists who cooperate with each other. One goes on to suppose that natural selection operates at the level of groups, so that communities with less cooperation will tend to perish. This response used to cause evolutionary theorists to roll their eyes, but apparently it has been gaining in plausibility. All that's required for group selection to operate is for the community to commit itself to punishing those who take advantage of their neighbors. This mechanism cancels out any gain from being naughty. Any armchair theorist can suggest this, but actually proving that group selection could work this way has taken a serious research effort over the last few decades. Dennett recapitulates it, especially the idea that to escape censure a member of a community must appear to be good, and that the easiest way to do that is to *be* good (although, as we all know, there are other ways that occasionally work quite nicely).

Chapter 8 is about what free decisions look like in the brain, and how easy it is to fall into the fallacious idea that there must be a single place and time where "will" operates. In his previous works, Dennett has ridiculed the "Cartesian theater" where the homunculus of consciousness observes the sensations brought to it by the senses; he now points out that there can't be a homunculus giving orders to the muscles either. I am under the strong illusion that my behavior follows uncaused decisions made by "me", but that doesn't mean that inspection of the brain will actually find a module where such decisions take place.

Making a decision requires work by the whole brain, over an interval of time that need not coincide with the time we believe the decision to have occurred.

The last two chapters of the book deal with the question, Can the concept of morality survive if everyone understands its evolutionary basis? One desideratum for a moral system in a democracy is that it be transparent, in the sense that its stability does not depend on concealing the truth about it from a large segment of the population. To give an example of "opacity", suppose the elite of a society decided that the way to maximize social harmony was to promulgate a certain religious system. They might be correct in their conclusion that a moral and just society would emerge if most people believed in the official religion, but this outcome would depend upon most people's believing that the religion was actually true, that its god maintained the moral balance in some way. Transparency would require that they know that *belief* in the god, not the fabricated god itself, was the mainspring of their society. But they couldn't be allowed to know this or the system would fall apart. A similar example is the society in Aldous Huxley's *Brave New World*, where people are programmed from birth to believe that being a member of their predestined caste is the best life to live. This system eliminates envy of castes above each citizen, and fear of castes below. But only the top group, the "alphas", know how the system works. The lower echelons really believe that a beta or a gamma is a wonderful thing to be, for reasons that seem convincing to them, even though the reasons are part of what was imprinted into their memories when they were born.

The fear that a scientifically sound theory of morality arouses is that, if everyone accepted it, they would see it as evolution (biological and cultural) causing everyone to be "glad I try to be good", just as in Huxley's world each character was trained to be "glad I'm a beta", or in general be glad they're whatever they are. Once people see the trick, their actual allegiance to being good might wither away.

Dennett's solution to this problem is to make accountability be the price one pays for freedom.[2] Even if a person has lost their faith in God, or lost the disposition to be good they acquired unquestioningly as a small child; even if the person is a psychopath with no moral feelings at all, they might still agree to be punished when they transgress, as a precondition to being readmitted to full citizenhood. If they're caught transgressing, they would submit to punishment as the only way to earn back their rights. The alternative would be to claim some disability such as insanity that would make the miscreant liable to close supervision for life.

This brief summary should make the book sound juicy enough; but there are many more topics than a reviewer has space to mention. If free will interests you at all, read the book; you won't be disappointed even if you disagree with Dennett. I basically agree with him, but there is room to quibble, so I will now quibble.

Is there any special connection between Chapters 1–4 of the book and Chapters 5–10? Traditionally it has been thought that the question of free will is tightly linked to the question of moral responsibility, but I think this is a mistake, for reasons I can only sketch here.[3] My first quibble is with the goal of Part 1 of *Freedom Evolves*.

[2] He acknowledges a debt to several other philosophers and social scientists for this idea.

[3] They are dealt with at greater length in [4].

One thing that is odd about the philosophical literature on free will is its preoccupation with determinism. We have known for several decades that the world is not deterministic, and one would expect this fact to cause philosophers to change the way they talk about physics and free will. It has, in fact, had no such effect, and in this Dennett follows tradition. He makes reference to quantum mechanics, but his prime "laboratory" for studying basic questions of freedom and causality is the two-dimensional world of the "Life game" cellular automaton [3], which is deterministic. Later, in Chapter 4, he explains how quantum indeterminacy wouldn't supply any help for those philosophers seeking the basis of free will in the natural world (and in so doing convincingly destroys those philosophers' attempts). But one can't deny that quantum indeterminacy is real, and pervasive. Many philosophers (I'm not sure about Dennett) believe that quantum effects are unimportant for macroscopic objects, and therefore are unimportant for us. The fallacy in this argument is that when physicists talk about macroscopic objects, they have billiard balls in mind. Billiard balls contain no interesting microscopic parts; every piece of a billiard ball is following whatever trajectory the ball does. But people have lots of microscopic parts, and it is perfectly plausible that quantum effects affect our behavior continually, normally in small ways, occasionally in larger ones.[4]

Perhaps Dennett focuses on the deterministic case because, if he can convince us that free will is possible in a deterministic world, our shock and awe will carry over to one that seems more open. I think this is a tactical error. We now know that the present state of the universe determines a probability distribution over its future states. A familiar staple of the literature of determinism is that in a deterministic universe all of Shakespeare's plays were present in the initial state of the universe, just waiting for time to reveal them. We now know that Shakespeare occupies a tiny subset of the ways time could have gone, and a particular play a tiny subset of the way the world could have gone after Shakespeare was born. Does this fact diminish the difficulty of the free-will problem? Not at all. Determinism is a red herring. With or without it, the problem of free will is explaining what a free decision looks like from the point of view of physics and biology.

It seems to me that the right way to answer that question comes in with the "choice machine" that I described earlier as a system that predicts alternative futures deriving from different actions and chooses the action that probably leads to the best future. It is reasonable to suppose that decision makers of this degree of complexity will have world models in which they themselves appear as entities. After all, their own bodies are things that show up wherever they go, and these bodies have special properties, inedibility for instance. The most important special property a complex decision maker must attribute to its body is that the body's behavior is governed by the decisions the decision maker makes, and therefore that it is completely useless to predict what that body will do without taking the decision into account. That means that, in the models the decision maker uses to predict the future, its body and the objects it can manipulate must be classified as decoupled from the causal laws governing other physical objects. Predictions of events affected by the agent's actions

[4] For one dramatic example, consider the effect of Brownian motion on sperm, and think about the sperm that started you off. It won the race against the other sperm, but how much help did it get from Brownian jostles? If we rewound the world's tape to a few minutes before your conception, what are the chances you would be born? Uncomfortably slim.

are always conditional on choices now being made. If you see your coat catch on fire, you do not try to predict what you will do, but instead to predict what will happen if you decide to take your coat off, or to roll on the ground, or to run toward the nearest river. So answering the question, "What does a free decision look like?" requires looking in a surprising place: at the decision maker's model of itself. A free decision looks like one whose outcome, according to the agent's self-model, is determined at least in part by the agent's deliberations.

I realize that this is a pretty deflationary account of free will. If it is correct, then free will is not a mystifying or mystical achievement, simply a minimal requirement for a decision maker that takes its own behavior into account. Someone born with a defective world model, lacking the normal view of events their actions can affect, would never be granted full membership in the human community. We would classify him as mentally ill in some way. Perhaps some versions of autism are due to such a defect. One of the most heart-breaking versions of autism is "childhood disintegrative disorder", in which a child develops normally for a couple of years, then goes into reverse, losing linguistic, social, and motor skills. To speculate a bit wildly, perhaps at just the age normal children are developing the ability to project alternative futures and choose a course of action, a child with this disorder sits around deferring decisions until he or she has come up with a prediction of what he or she will do. The poor thing looks for causal regularities to explain its bodily motions, and soon enough finds them: as time goes by, its body departs from purposeless activity less and less, and becomes more and more predictable. Any tendency to make a choice would decrease predictability, and would tend to be extinguished over time.

I sketch this possibility just to make it clear that free will is nothing but one of the basic tickets for entry into being human, and one of the least complex at that, not nearly as impressive as, say, the ability to learn syntax, another basic ticket. (Actually, while syntax is specifically human, the free-will ticket must get punched for many complex animals.)

If this is correct, as I believe it is, then what does free will have to do with moral choice? The usual link between them is this argument: someone should be punished for an action only if he or she could have taken a different action. Suppose someone commits a prima facie criminal act, but turns out to have had no real choice; then it would be immoral to punish her. If our decisions only appear to us to be free, but are not "truly free" because they are caused, then she had no real choice—she could not have acted any other way—and we should not punish her.

I am not the first[5] to point out that this argument is incoherent, not for the reasons Dennett cites, but because it mixes up two conceptual frameworks: one in which we do make free choices, and one in which we are stuck in the causal flow. If we (whoever "we" are) are discussing what we should do about our prisoner, we use framework F (for freedom); if we are predicting or explaining someone's behavior, we use framework C (for causality). It is not always necessary to keep these strictly apart. If some psychologists are studying the behavior of an organism, they will keep their own options open ("What experiment should we do next?"), while ignoring what the test subjects think about their options. No contradiction arises even if the subjects are the same species as the psychologists. Advertising

[5] Immanuel Kant or William James might lay claim to be the first.

executives draw similar lines. The problem is that in the ethical argument human agents are assumed in the same breath to be entirely subject to causality (frame C) and entirely free (frame F), because the ethical argument is meant to cover *all human decisions*, including the one we are about to make about how to punish our prisoner. To avoid inconsistency, we have to split it into two arguments:

A_C: The probability that the prisoner would choose to commit a crime was completely set at the time she chose to commit it. The operation of her decision-making facility was governed entirely by that probability distribution. The decision we are making about how to punish her is governed by an analogous probability distribution, which we do not and cannot know.

A_F: The prisoner knew the harm she was doing to others, and the consequences if she were caught. She chose to commit the crime anyway; no one coerced her or tricked her. However, it was her first offense, so we should give one year of probation.

Each argument is unproblematic. The second is, of course, familiar, because we all engage in that kind of reasoning all the time. The first is a bit unusual and not terribly useful, but it's correct as far as it goes.

Some might say that argument A_F is "ungrounded", in the sense that, like all arguments about what to do, it presupposes at every step the truth of the delusion that human decisions are exempt from causal laws. But the delusion is inescapable and universal. If it is possible to make a good argument within framework F, and if everyone we talk to lives within the same framework, then there is no good reason to require us to prove that the argument is sound from the viewpoint of another framework. By analogy, constructivists reject the proofs of mainstream mathematicians as based on delusions such as the law of the excluded middle. But even the constructivists grant that there is a distinction between a valid non-constructivist proof and an invalid one, and the former are not those that have independent constructivist support.

Please don't take my position as a variety of *relativism*, the doctrine that different groups can have different notions of truth, and that it is meaningless to talk about a proposition as objectively true or false. Framework C is obviously correct, and most conclusions drawn within framework F are simply gibberish (not even false). Knowing that, should we reject those conclusions? If you're tempted to answer this question one way or another, you are still in the grip of framework F. Welcome back to the asylum. If we have to choose between being mindless or being deluded about how causality affects us, the choice is obvious.[6]

From this point of view, some of what Dennett says in Part 2 of his book seems unnecessary. The title of the last section of the last chapter is "Human Freedom is Fragile", which is misleading. Of course *political* freedom is fragile, and eternal vigilance is, as they say, its price. But *metaphysical* freedom is not fragile, or at least fragile only in the sense that humanity is fragile. As long as we survive, and do not degenerate into marmosets, we will have free will. Recall that one purpose of Chapters 9 and 10 is to allay fears that knowing the basis of free will, that is, having a "transparent" ethical system in the sense that everyone knows how it works and how it came to be, would cause people to stop behaving ethically, or degrade their desire for political freedom. His

[6] "We have to choose??" the framework-C fanatic pleads, "Can't we find some other way to phrase that?"

argument that this is unlikely, which I summarized above, is entirely convincing. But there is a simpler argument, which is that knowing what your brain does when you do X can have zero impact on how you do X. To cite an example of Jerry Fodor's, a foreign language, in contrast to our own, sounds like a meaningless sequence of noises. If we know a few words of the language, one of them will pop out of the babble every now and then, leading us to the implausible claim that people talk faster in that language than we do. Our own language, of course, doesn't sound like a sequence of noises at all. It sounds like a sequence of meaningful sentences, accompanied by extra labels regarding intonation, accent, pitch, gender, and sometimes the identity, of the speaker. ("Ironic, Boston accent, high pitched, female, my wife".) But we know that the sounds reaching our ears are of exactly the same genus as the sounds of a foreign language, and that before phonological, syntactic, and semantic processing they are the same sort of babble. We talk fast, too. Knowing that, listen to a fellow speaker of your language and ignore the words and sense, just attend to the babble that you are sure is what reaches your ears. You can't. In Fodor's [2] rhetorical question, addressed to a monolingual English speaker: you know what French sounds like; what does *English* sound like? Similarly, knowing that free will is just a standard fix to a bug in the idea of predicting the future, try making a decision without assuming your options are open. You can't.

Similarly, if a group of humans wants political freedom enough, they will fight for it, paying no heed whatsoever to the fact that their decision to fight is made by mechanisms subject to physical laws. The real danger is, as it always has been, that people will have no chance to win political freedom, or rate it as less valuable than things like food, or safety from terrorists. We don't need philosophy to tell us this.

On the other hand, it seems to me that on the really deep moral questions, Dennett's account falls short. It explains why the "ethics game" exists and why we all play it. But some important moral decisions are beyond the scope of the game. Robert McNamara, in recent reminiscences,[7] says that he and colleagues planning the firebombing of Japanese cities in World War II were aware that if the US lost they could be prosecuted as war criminals. But it's unlikely that they thought losing was a serious possibility; they weren't trying to get away with something that they would have to atone for later if caught. From their point of view, they judged, correctly, that no one would know or care exactly what they had done, so long as it contributed to victory. What do we say about such a case? I don't mean to imply that the answer, or even the question, is clear. I just want to point out that evolutionary theory doesn't give us any guidance. It doesn't say much of anything, except perhaps that there was selective pressure against living in cities during World War II, and that if we keep fighting wars like that, evolution may favor people who prefer small villages, or perhaps people who are more like marmosets.

My conclusion is that, while a scientifically sound conception of morality doesn't threaten the foundations of morality, it doesn't do much to support them either. For that we have to fall back on traditional considerations, and talk about traditional concepts such

[7] From Errol Morris's film *The Fog of War*, as quoted at http://www.alternet.org/story.html?StoryID=17508.

as categorical imperatives, the greatest good for the greatest number, and maybe even God.

Quibbles and blemishes aside, *Freedom Evolves* is another winner from Daniel Dennett. If you have the slightest interest in free will, evolution, and morality, you will be amply rewarded by reading it. Even if it doesn't convince you, it may force you to find new justifications for some beliefs about human freedom that you assumed could not be challenged.

References

[1] G. Drescher, Made-up Minds: A Constructivist Approach to Artificial Intelligence, MIT Press, 1991.
[2] J. Fodor, The Modularity of Mind, MIT Press, 1983.
[3] M. Gardner, Wheels, Life, and other Mathematical Amusements, W.H. Freeman and Company, 1983.
[4] D. McDermott, Mind and Mechanism, MIT Press, 2001.
[5] C. Taylor, D.C. Dennett, Who's afraid of determinism? Rethinking causes and possibilities, in: R. Kane (Ed.), Oxford Handbook of Free Will, Oxford University Press, 2001.