Max Weber Lecture Series

MWP – LS 2011/01
MAX WEBER PROGRAMME

""MY BRAIN MADE ME DO IT" (WHEN NEUROSCIENTISTS THINK THEY CAN DO PHILOSOPHY)

Daniel C. Dennett

EUROPEAN UNIVERSITY INSTITUTE, FLORENCE
MAX WEBER PROGRAMME

"My brain made me do it"
(when neuroscientists think they can do philosophy)

DANIEL C. DENNETT

MAX WEBER LECTURE No. 2011/01

ISSN 1830-7736

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
www.eui.eu
cadmus.eui.eu

Abstract
Some philosophers and neuroscientists have recently been saying that science shows that we don't have free will, but it turns out that this claim−which would be bad news if true−is due to misrepresentation and misinterpretation. Since free will matters to people, and should matter, these contributions to public misunderstanding are regrettable. When we clarify the issues we see that we will have to make some significant adjustments to our understanding of moral responsibility, allowing for more differences in moral competence than our traditional understanding recognizes.

Keywords
free will, cognitive neuroscience, Libet, determinism, randomness, consequentialism, retributivism

The lecture was delivered on 15 December 2010

Daniel C. Dennett
Center for Cognitive Studies,Tufts University

Are philosophers harmless? I think that the presumption is that we are a pretty harmless lot. For instance, ask yourself: would any philosopher ever need to get malpractice insurance? What could we do? We seem to be fairly ineffectual, but I think actually this stereotype should not be accepted. Sometimes philosophers can do some serious mischief, and I think it is time for philosophers to start emulating engineers and other scientists, studying the impact of their work on the conceptual and social environment. That is where we can do some serious damage and I am going to be talking about one such prospect today. We have made a mess of the free will issue, and we have been working at it for several thousand years. And now the damage we have done is being amplified by neuroscientists and lawyers. When I composed the title of my talk, I was feeling fairly grumpy about the neuroscientists, but the more I thought about it, the more I thought that the philosophers are really the ones who should be blamed for the problems that we are facing. The neuroscientists are just tagging along and taking some of what we have said more seriously than they should. The result is a rather unfortunate stew of misapprehensions, which I'm going to try to separate and clarify.

"My brain made me do it." I keep running into this phrase in various places. Is it, as some have surmised, a usable defense in a criminal trial, for instance? "I couldn't help it; my brain made me do it!" There is at least one book by that title, and also an article by a distinguished psychologist who is, I am happy to say, dismissive of the ideas evoked by the phrase—but not dismissive enough (Bloom, 2008). What is puzzling about this phrase can be brought out by comparing it with "my *mind* made me do it," which seems on the face of it to be an admission of personal responsibility. "No, I wasn't pushed; it wasn't an accident; I decided right there and then—my mind made me do it." But since your mind *is* your brain what else would you want to make you do it, if not your brain?

What we must understand, of course, is that your brain must make you do it in the *right way* and then everything is fine, but what is that? What is the right way? How is there or could there be, in fact, a right way that our brains could make us do the things that we do so that we would be responsible for them? Does the right way require *indeterminism*, as some people think? Tradition says yes. There is a tradition of people thinking:

> 1) I cannot be responsible without free-will (that is "by definition"), and
> 2) I cannot have free will if my decisions are all physically determined (that too seems true by definition to many people).
> 3) Therefore I cannot be responsible if determinism is true.

This seems like an obviously sound argument until we raise the prospect that "free will" is being understood in different ways in the two premises. This is an avenue well worth exploring, since determinism may well be the truth about our brain-caused actions, and if people do not believe they have free will, they will tend to conclude that responsibility is a myth, . . . and then maybe they will start behaving badly. This has long been the worry of some philosophers: if people get the idea that moral responsibility is just a myth, and that science is exploding that myth, then they will stop trying to lead moral lives. They will not take their own morality seriously anymore. You might well think: only a philosopher would really worry about that. But no, there's actually some new and disturbing evidence that this is the case. Here is a passage by Francis Crick from his book The Astonishing Hypothesis:

> You, your joys and your sorrows, your memories and ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules. Who you are is nothing but a pack of neurons. So although we appear to have free will, in fact, our choices have already been predetermined for us and we cannot change that.

This passage was read by a group of subjects in a pioneering experiment by Vohs and Schooler (2008). They had subjects read this passage and they were then given a puzzle to solve, and paid depending on whether they got the solution right. The experiments gave them an apparent opportunity to cheat without being detected. The control group read a different passage from the same book, but it wasn't about free will at all. And yes, those that read the quoted passage about the myth of free will cheated significantly more than those who read the control passage. This is not a lone result. It has been followed up by further research, both replications and studies with different experimental conditions. When subjects are first presented one way or another with the idea that they don't really have free will, they do tend to behave less morally than otherwise.

So neuroscience—if the Crick passage is taken as reporting a discovery of neuroscience—does seem to threaten free will; it seems to show, in fact, that we don't have any. Scott Adams has presented the issues with customary vividness in his cartoon strip, Dilbert:

> Dogbert: Do you think the chemistry of the brain controls what people do?
> Dilbert: Of course.
> Dogbert: Then how can we blame people for their actions?
> Dilbert: Because people have free will to do as they choose.
> Dogbert: Are you saying that free will is not part of the brain?
> Dilbert: Of course it is, but it's the part of the brain that's out there just being kind of free.
> Dogbert: So you're saying the free will part of the brain is exempt from the natural laws of physics?
> Dilbert: Obviously, otherwise we couldn't blame people for anything they do.
> Dogbert: Do you think the free will part of the brain is attached, or does it just float nearby?
> Dilbert: Shut up.

This actually captures the core problem very well. Paula Droege says, in a recent article: "Recently, I attended a lecture by an eminent neuropsychologist [Patrick Haggard], who declared that neuroscientists have to believe that conscious intention is an illusion." So this is an idea that is out there in the *Zeitgeist* from some very eminent and influential scientists who are not shy about talking to the general public. I am all for them talking to the general public, I just think they sometimes should get a little more advice before they do it.

The traditional philosophical quandary is the conflict between determinism and free will. Determinism is the idea that every event has a cause, which has a cause, which has a cause, in a causal chain that goes back to the Big Bang, if you like, and that there are no events without causes—*undetermined* events. That is the traditional quandary, and it is perfectly expressed by Dilbert, but it is, I believe, a red herring. There is a more important and troubling issue, which is how to undo the misunderstandings of the implications from neuroscience—with or without determinism—for human responsibility. I have just said that determinism is a red herring, but I cannot expect you just to take my word for it since there is a 2000 year-old tradition, even older than, in fact, to the contrary. So I am going to have to say a little bit about that. I will first spend a little time supporting the idea that *determinism vs indeterminism* is a red herring, then I will look at the question "does neuroscience show that we don't have conscious free will?" and finally I will look more constructively at what *would* give us the free will. Then we will be able to see whether we have free will or not.

Indeterminism is a red herring.

I need a spokesperson, and when I need a spokesperson I often go to my old friend, Jerry Fodor, who can be counted on to say something vivid, unforgettable—and false. Jerry is a sort of human trampoline; often it is the case that if I can see farther than others, it is because I am jumping on Jerry. (While I am paraphrasing that famous remark of Isaac Newton's, I have to add that I regret to say I was not the author, although I was once quoted as the author, of another variation on that line; somebody quoted me as having said, "If I can't see as far as others, it's because giants are standing on my shoulders!" I wish I'd said it, and I've said it now, but alas I am not the author of it.)

So what does Jerry Fodor say about what we want? He says:

> One wants to be what tradition has it, what Eve was when she bit the apple, perfectly free to do otherwise, so perfectly free in fact that even God couldn't tell which way she'd jump. (Fodor, 2003)

In other words, what one wants is a miracle. Or magic. There is a wonderful book by Lee Siegel, *Net of Magic: Wonders and Deceptions in India* (1991), about the history of Indian street magic, the source of much if not all of the rituals and adornments of stage magic. Siegel himself is a philosopher, novelist and magician, and I highly recommend his book, both for what it reveals about the performance of magic in general, but also for its insightful perspective on Indian customs and beliefs. There is a passage in that book which I have very much taken to heart. Indeed it has become a sort of talisman for me. He says:

> I'm writing a book on magic, I explain, and I'm asked, "real magic?" By *real magic*, people mean miracles, thaumaturgical acts and supernatural powers. "No," I answer, "Conjuring tricks, not real magic." *Real magic*, in other words, refers to the magic that is not real, while the magic that is real, that can actually be done, is *not real magic*. (p425)

This nicely sums up a problem that I've been confronting all my career, and on two different issues: free will and consciousness. For many people, if your theory of consciousness does not have it coming out to be *real magic* then whatever you are talking about, it is not consciousness. You are just explaining it away rather than explaining it. And it is the same with free will: if you come up with a version of free will that is not *real magic* then you are just not talking about the free will that they are interested in. And so then the task becomes, how do you get people to trade in their inflated desire for real magic for an appreciation of the wonderful conjuring tricks that evolution and nature have given us that really do the job that needs to be done? A lot of people just do not like that bargain, and will not even consider it. There is a tradition of this, going back several hundred years at least. No less a philosopher than Immanuel Kant famously dismissed the view that I will be defending here, a version of *compatibilism*, as a "wretched subterfuge." He wanted real magic for his view of free will and would accept no substitutes. Fodor and Kant are not alone, here is another philosopher, Galen Strawson, in another review of my book: "He doesn't establish the kind of absolute free will and moral responsibility that most people want to believe in and do believe in; that can't be done and he knows it." (*New York Times,* 2003). Exactly right. That cannot be done, and I know it. I cannot give people the kind of absolute freedom and responsibility that they want and I do not even try. I tell them they will have to accept something a little less magical, but still good enough. What neither Fodor nor Strawson, or most of the other people of that persuasion, do not even try to do is to *defend* this common folk desire. I agree with them that that is what "one wants," that is what most people want, that is even what common sense says you should want. And I say they are all just wrong, It is an indefensible desire, however common and natural it is; it does not get you what you think it does and so your desire for that kind of freedom is actually a rather deep mistake. In fact, free will has nothing to do with indeterminism. Well, *almost* nothing, as we shall see. I am going to keep this part of my talk short, since I and others have done justice to it at great length, and because I want to get to some more novel issues, but let me give you a little thought experiment to give you a sense of why indeterminism is just not the issue.

This is going to be a *reductio ad absurdum*, so *suppose* that free will *did* depend on indeterminism, but that your particular source of indeterminism was something that you carried around with you—such as a box containing some radium randomly emitting radiation, and a Geiger counter to detect it. You are about to get on an airplane with your handy randomizer, but security will not let you take your randomizer along. Heavens! How will you be able to make responsible free choices without your handy portable source of genuine indeterminism? You plead with the security people but they are adamant. Are you doomed to a life trajectory without free choices until you get home? Not at all. Here is what you can do. Before your flight you go off to the rest room (for privacy) with your randomness generator, and get your randomness generator to spew out a few dozen or a few hundred genuinely random numbers, which you write down, in secret, on a piece of paper you put in your wallet. (How many do you think you will need? Be sure to take along enough to last you till you get

back!) (If you think the flip of a fair coin is random enough for you, then just do a few hundred secret coin flips and carefully write down the results: heads, heads, tails, heads, tails, tails, heads . . .). In any event, by recording a series of genuinely random results, you then have a handy source of randomness that you *can* take on the plane. And it will work *exactly as well as* if you had the randomizer at your side generating randomness on a just-in-time basis, to give you the unpredictability you crave. Any time on your trip that you need a random number, you just look at the list, use the next one on the list, cross it off and keep going. Whatever benefit you could have got in your decision-making from having a portable radium randomizer, even a source of quantum randomness locked up in your brain, you can get just as well from the list you carry around with you.

In other words, if you need indeterminism for your free and responsible decisions, you can get it from undetermined events occurring in your brain at decision time, *or* from undetermined events that happened long before, and have just taken their time getting into position to play their role in your decision-making. Whatever effect an undetermined event could have by occurring in your brain at noon today could also be produced by an undetermined event that happened long before you were born, far away, beamed deterministically at the speed of light to your brain, and arriving at noon.

The list in your pocket is as good as any real-time randomness generator in your brain. Well, *almost* as good. There is one exceptional condition—and one only, so far as I can see—where indeterminism *could* make a difference that mattered. It is worth our attention because it helps me prove a different point, not directly about free will, but about something that does matter. If you have enemies or competitors who are particularly inquisitive, who might be able to get a peak at your list, then, and only then, would it behoove you to get your random numbers ready-made, on the spot, so, as Fodor says, "even God couldn't tell" which number would come out. Then there is simply no way for anyone to read the list. Here is an interesting fact: all computers have so-called random number generators in them, because many programs use random numbers as tie-breakers, when the program has to do *something* and has no information about which option to start with. For instance, the program must find an item that has a certain property and is presented with *n* candidates. Which to test first? It gets on with it by picking a candidate *at random* and testing it first, continuing on until it finds one that passes the test. It consults the random number generator to get a coin flip, a roll of the dice, to help it over its indecision. Those numbers are not really random, however; they are pseudo-random, generated by a process that could, in principle, be reverse-engineered; this is equivalent to having the list in your pocket that could, in principle, be peaked at by an enemy. It is possible to purchase a genuine quantum randomizer that closes off this possibility. Who would pay for such security? Cryptographers. Because they make their living in a world where there are competitors, enemies who are seeking by all possible avenues to get the list, and if they got the list, then they would be able to break your codes. Cryptographers have a use for genuine, right-here-and-now randomness because they are rightly concerned about not having their minds read.

Here is a simpler example: the game of rock-paper-scissors. It is provable that the best strategy, the only unbeatable strategy, is to play randomly. Then there simply is no pattern that can be discerned by an opponent and exploited. That way, your opponent cannot track your moves at all. But when you play, you do not want to consult the list in your pocket too early, because then you might involuntarily reveal your choice to your opponent. As a poker player would say, you want to avoid having any "tell" – something about your facial expressions or your manner that telegraphs what you're going to do. If you have a tell, a really clever and quick opponent is going to pick up on that and take all your money from you. So if you are playing rock-paper-scissors for money, hide your list of random numbers, do not let anybody see it, and do not even let *yourself* see it until the last split second. Being a cryptic chooser can be important in the dog-eat-dog world we live in, and if you are playing rock-paper-scissors *with God*, you had better play for genuine randomness, just as Fodor says. Otherwise, the cheaper substitute is just fine, as long as you guard your brain from snoops.

But still, you may think, in a deterministic world, there are not any *real* options, any real opportunities. A favored image, not because it is good, but because it nicely captures the intuition that people have is this: you go to Disneyland and go on the jungle boat ride, and the captain makes a dramatic show of *almost* steering into one catastrophe after another. Oh, you just missed that big hippopotamus, and then the crocodile, and—close call!—you narrowly avoided the waterfall! But of

course the boat's helm is completely non-functional, since the boat is running on an underwater track. There was never any chance it was going to go off the hidden rails and collide with anything. There was a delicious illusion that there were these opportunities for disasters, all of which were deftly prevented from happening by the quick work of your skilled captain. A lot of folks think that this is the true face of determinism. "If determinism is true, my whole life is sort of on these hidden railroad tracks, and I don't actually have any free will at all!" But, in fact, there is a fundamental difference: the activities, the desires, the reflections of the helmsman, the captain on the jungle boat are *causally inert*; they are not playing a role in determining the trajectory. But when *you* make a decision, the reflections going through your mind *are* in fact playing a role. Which is just what you should want.

Suppose you are playing baseball and it is the bottom of the ninth inning with two outs, the bases are loaded, the game is tied—and you are a bad batter! Here comes a pitch. If you just let the pitch hit you then you are awarded first base and that forces in a run in and your team wins. But it is going to hurt, so your natural tendency is to duck. If you duck you are going to *avoid* being hit by the pitch, but understanding the situation, reflecting on it, you *could* decide, "this time I'm going to thwart my reflexes and I'm going to let the pitch hit me." You can *avoid* avoiding the pitch. But maybe a fan of the other team has bribed you not to do that, so you think, "No, in this case I will thwart my desire to thwart my wish" and you thereupon avoid avoiding avoiding the pitch. The fact is that we *can* reflect on these matters in real time, and when we do, the reflections determine what we do. That is the difference between us and the jungle boat captain. The fact that these reflections themselves have causal antecedents does not make them any less effective.

One more poke at the red herring: People say "You can't change the past, but you *can* change the future." It seems right, doesn't it? But from what to what? You cannot *change* the future. So suppose somebody says, "If determinism is true, I can't change the future." That is true. If determinism is *false* you cannot change the future either! From which it follows: you cannot change the future. So forget about it, it is not in the cards; and besides, it is not what you want to do. What you want to do is to bring it about that the future that happens is not the one that would have happened if you had not acted. Thus, here comes that pitch; it is *going* to hit you. You duck and the pitch goes over your head. Have you changed the future? No because as it turns out, the pitch *was not* going to hit you because you saw it in time and ducked. That is what we want: to be able to rise to opportunities of that sort and, in fact, in that situation determinism actually helps you because it gives you nice law-like regularities. If you want to avoid being hit by things, it is better that they be baseballs than lightning bolts, because to some degree you can predict where baseballs are going to be. That prediction depends on the deterministic nature of the baseball's trajectory. The philosopher David Wiggins wrote some years ago eloquently deploring what he called the "cosmic unfairness of determinism" (1973, p.54). Why didn't he also talk about the cosmic unfairness of *in*determinism? If indeterminism is true you are just as much victimized by the random events that influence your behavior as, in a deterministic world, you are influenced by the *non*-random events. There is really no difference here; the sense that there is an asymmetry here is just a mistake. The truth is that you cannot control everything, whether determinism is true or indeterminism is true, but you can control some things and that is what matters. We do not need absolute free will; we just need pretty good free will, which we can have. Here ends the first part of the talk.

Does neuroscience show that we do not have conscious free-will?

An influential article by a philosopher and a psychologist Josh Green and Jonathan Cohen, "For the law, neuroscience changes everything and nothing," has been much cited since it appeared in *Philosophical Transactions of the Royal Society* (2004).

> The law says it presupposes nothing; the law says that it presupposes nothing more than a metaphysically modest notion of free will that's perfectly compatible with determinism. However, we argue that the laws' intuitive support is ultimately grounded in a metaphysically over-ambitious libertarian [by *libertarian*, they mean an indeterminist notion of free will] that is threatened by determinism and more pointedly by forthcoming cognitive neuroscience. (p1776)

They go on:

> New neuroscience will undermine people's common sense, libertarian conception of free will and the retributivist thinking that depends on it, both of which have heretofore been shielded by the inaccessibility of sophisticated thinking about the mind and its neural basis.

So they are anticipating that neuroscience is going to undermine everyday, common sense and this libertarian, traditional, indeterminist notion of free will.

> Free will, as we ordinarily understand it, is an illusion generated by our cognitive structure. Retributivist notions of free will ultimately depend on this illusion and, if we are lucky, they will give way to consequentialist ones, thus radically transforming our approach to criminal justice.

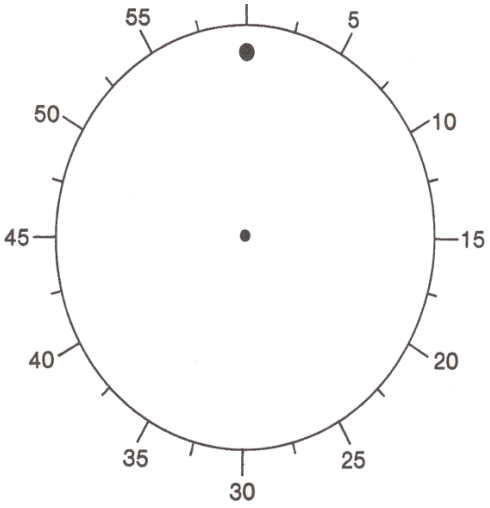
There is a lot going on in this passage. First of all, let me remind you that retributivist accounts of punishment, such as Kant's, say that the point of punishment is not just to rehabilitate, or to deter other evildoers; punishment is a good in itself. Kant infamously said that if the world were going to end tomorrow, one of our duties in the last few hours of life on earth would be to execute the people on death row, because the world would be better for our having punished them for their crimes. That is what you might call industrial strength retributivism. Consequentialism, as the name reminds us, holds that punishment, like any other response to misdeeds or antisocial behavior, is only justified in virtue of its consequences for the future. Green and Cohen say if we're lucky, the effect of the new neuroscience will be to erode popular support for retributivism and replace it with a consequentialist vision, which they think would be radical—but a good thing. They start off by saying, "Free will, as we ordinarily understand it, is an illusion generated by the cognitive structure." Well, yes, as we *ordinarily* understand it, I think in a way that is right. The ordinary, common sense view, the view that Fodor articulates as does Strawson, is in fact an illusion generated by our cognitive structure, but there is a non-illusory view of free will which is available. So what I want to do is replace the second half of their statement and say, "If we improve our understanding of the ordinary concept of free will, we can split the difference between retributivist and consequentialist notions of punishment, thus gently and humanely transforming our approach to criminal justice." We do not want to get rid of retributivism root and branch, we just want to tame it and reform it. I will say why. To a first approximation, the reason has been well known for many years. One of the most vivid portrayals of it is Stanley Kubrick's movie, *A Clockwork Orange,* and the novel it came from by Anthony Burgess. Burgess was once asked in an interview for the significance of the title, and he said, "An organic entity, full of juice, sweetness and agreeable odor, being turned into an automaton." This is the anxiety-producing image he wanted to suggest, and it is very much the sort of image people in the neurosciences these days are conjuring up in various ways. I am no stranger to it. In fact some years ago Giulio Giorello, a fine journalist and philosopher of science, interviewed me in the *Corriere della Serra* (Milan, 1997) ; the headline of the published article was ***Sì, abbiamo un anima. Ma è fatta di tanti piccoli robot.***

"Yes, we have a soul, but it's made of lots of tiny robots." Exactly right. This has become my slogan because it perfectly expresses my view. We do have a soul—whatever it is that gives us free will and responsibility, that makes us moral agents—but it's not an immaterial, immortal soul, it is a structure in our brains made of lots of tiny robots. Our bodies are approximately 100 trillion robotic cells and nothing else. It is the teamwork of those cells working together in ways that are trained, governed, inspired, adjusted, modulated by the culture we stuff into our heads: that is what gives us a soul. Our brains *are* clockwork oranges: there is no wonder tissue, there is no immaterial soul. Here is a quote that bears on the topic, "We thought we were special and reductionism seemingly shows that we are not; we too are just machines." (Moore, 2011) Just machines. Well, so what? Machines can be pretty wonderful. Why say "just?" We are not special in the way that tradition would have it – a little below the angels – but we are still pretty special because we are very special machines. But this is what is being challenged by some work in neuroscience.

The most important work, the work that is most influential, is work done many years ago by Benjamin Libet (1985), and some recent work done by Daniel Wegner (2002), but I am not going to

discuss Wegner today (I have discussed his work at length in Dennett, 2003)**,** since I want to concentrate on disarming Libet's experiment on voluntary action.

In fact we are going to recreate his experiment informally right here, with you as subjects, so you can see what it was like. First, we need a very simple voluntary action that can be easily recorded in time, and I want you to practice now. Sometime in the next ten seconds or so, I want you just to flick your index finger, either left or right index finger; just flick it please. Not so difficult, was it? Now, I have to make your task a little more complicated: you are not to *plan* this flick; you are not to do it for any *reason*; you are supposed to flick spontaneously for no reason at all; this is to be an *acte gratuit*; you are supposed to just *let it happen* whenever—dare I say—the spirit moves you. So give it a try one more time. Now I am going to complicate your task just a little bit more. Do not worry, I am not going to bring out the electrodes used by Libet in the experiment. Subjects had non-invasive surface electrodes placed on their scalps attached to an EEG machine, to record the electrical activity in their brains. You can pretend you have electrodes on your scalps, if you like. But now here is the tricky part: subjects were to look at this clock face (animated in the PowerPoint slide) with a little spot that goes round and round and round, making a circuit in about three seconds.



Now this time, when you flick your finger I want you to *notice* the position of the black spot at the very instant you become conscious of your urge or decision to flick your finger. . . . . Done? Let's just see how people did. For how many of you was the spot between the 5 and the 10? For how many, between the 10 and the 20? For how many, between the 20 and the 30? 30 and 40? 40 and 50? Okay, you get the idea. But now, I want to know how many of you found you were doing the following thing: you (sort of) decided, "Well I guess I'll flick when it's at the 15." Or at some other number? How many of you did that? [quite a few hands are raised] All right, those with the hands up, your trials were pre-planned; they do not count. Throw them out. You were not following instructions. If you decide to time your flick to a particular location of the clock that is not obeying the instructions. You are instead *simply to see* where the spot is when the *spontaneous* flick happens. I think you can now see this is a rather weird experiment. Your action is not much like the free will choices or the sorts of decisions that matter to us. This is a point that is often made, but I wanted to drive home just how strange this is before turning to what Libet found.

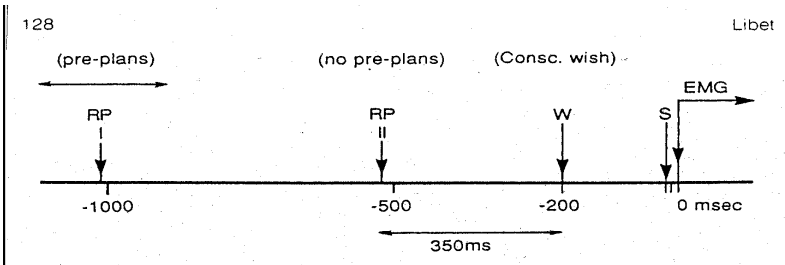


FIG. 3. Diagram of sequence of events, cerebral and subjective, that precede a fully self-initiated voluntary act. Relative to 0 time, detected in the electromyogram (EMG) of the suddenly activated muscle, the readiness potential (RP) (an indicator of related cerebral neuronal activities) begins first, at about − 1050 ms when some pre-planning is reported (RP I) or about − 550 ms with spontaneous acts lacking immediate pre-planning (RP II). Subjective awareness of the wish to move (W) appears at about − 200 ms, some 350 ms after onset even of RP II but well before the act (EMG). Subjective timings reported for awareness of the randomly delivered S (skin) stimulus average about − 50 ms relative to actual delivery time.

The interval between RP and W is the famous Libet gap. But notice, on the left, the timing of the "pre-plans" trials; that's the average of the trials where subjects disobeyed. Ignore them. Here is how Libet got this data. After each trial, just as I did with you, he asked subjects to say in their own good time what the position of the clock face was when they were conscious of deciding to flick. The clock face was synchronized with the recording of the brain waves, so for each trial he could mark on the timeline what time it was (relative to the EEG) when you became conscious of the urge to flick. And what he discovered was that some 350 milliseconds before that moment, on average, there was a spike, called the "readiness potential," which is readily detectable in the brain and that predicts that you are going to flick. Now, remember not to look at the 1000 millisecond anticipations on the left, because those are the cases where subjects did preplanning. The 350 millisecond gap represents the average time—a third of a second—between the onset of the readiness potential and the subjectively (and retrospectively) reported onset of consciousness of decision, at time W. Notice that from W, the conscious urge, to execution takes 150 ms or thereabouts.

Here is what some eminent neuroscientists have said about it, Michael Gazzaniga (1998): "Libet determined that brain potentials are firing 350 ms before you have the conscious intention to act, so before you are aware that you are thinking about moving your arm, your brain is at work preparing to make that movement." William Calvin (1990): "My fellow neurophysiologist, Ben Libet, has, to everyone's consternation, shown that the brain activity associated with the preparation of movement, something called the readiness potential, starts a quarter of a second before you report having decided to move. You just weren't yet conscious of your decision to move, but it was indeed underway." (p80-81) And here is Libet himself: "The initiation of a freely voluntary act appears to begin in the brain unconsciously, well before the person consciously knows he wants to act. Is there then any role for conscious will in the performance of a voluntary act? To answer this it must be recognized that conscious will [that is the W] does appear about 150 ms before the muscle is activated even though it follows the onset of the readiness potential." (1999, p75)This gives rise to what is known as the Libet veto window, you have 150 ms from the time you are first conscious of your urge in which you can veto your urge. You can say, "Nope, I'm not going to do it." That is a little bit more than a tenth of a second, not much time, but that is a window of opportunity for issuing a veto on this. As Libet says, "An interval of 150 ms would allow enough time in which the conscious function might affect the final outcome of the volitional process." This vision of what Libet discovered led to an amusing remark by V. S. Ramachandran (1998): "This suggests our conscious minds may not have free will but rather free won't." (p35) It is a cute joke, but why would that be free? Why would the conscious function be free? After all, wouldn't there be earlier brain events churning along to determine whether the conscious function vetoed or not?

This, by the way, reveals a defect in Libet's study that escaped my notice until it was recently pointed out to me by Bob Doyle (personal communication): the only data Libet shows us are the averages of times in those cases where subjects *did not* avail themselves of the veto opportunity; the

data are put into registration, that is, by the flick that ends a trial. So we have no data at all on how often there are RPs that turn out to be false alarms because of vetoes. So we really cannot say whether the RP, the readiness potential, is a good predictor of subsequent flicking. Libet speaks to the issue of whether or not the conscious function can itself be unfree because determined by earlier events. He says, "The possibility is not excluded that the factors on which the decision to veto is based, do develop by unconscious processes that precede the veto." If this possibility is not excluded, then Libet has no evidence of a minor role for "free" conscious will, and hence this parenthetical acknowledgement by itself cancels the standardly received implications of the experiment. But, although this sort of critique has been issued many times, people just cannot get out of their heads the idea that there is something ominous or radical about this particular experiment. Sometimes, it turns out, there is a readiness potential that precedes by several hundred milliseconds the conscious event identified by subjects as their decision to flick "spontaneously". Why should this surprise us or upset us? When I asked you to perform this stunt several minutes ago, you did your best to comply (and to avoid pre-planning) but then what did you do? You more or less told yourself not to plan, and waited, in a state of alertness, for something to happen. Something did happen, and bingo! –that was your choice. Now you have no introspective access to the process that you somehow set in motion to get this to happen (if you did, you'd be pre-planning). So perhaps we can say that in normal subjects asked to do this, a process is initiated that takes several hundred milliseconds to mature into a flick. This does not shed any light, and in particular any ominous light, on whether we have free will in any sense that matters.

The worry raised by Libet makes sense only to someone who presupposes that consciousness, or the conscious self, is somehow apart from and distinct from the rest of the brain. That is the perspective that yields the worry: "Uh oh, am I out of the loop? I, my conscious self, may be somewhere in the brain, but out of the loop where the decision-making happens!" This vision of there being this place in the brain where consciousness is, is itself incoherent, the idea I call the Cartesian Theater, the place in the brain where it all comes together for consciousness (Dennett, 1991). When I go on the warpath against the Cartesian Theater—I have been doing that for 20 years and more—people sometimes say, "Well wait a minute. Are you making an empirical point or a conceptual point?" And my answer is I am making both. The conceptual point is that at some point we have to get rid of the Cartesian Theater, but it is an empirical fact that we have to get rid of it at step one. It is always possible, it is conceivable, that there could be a Cartesian theater. One, or even two. There is a very clear, and coherent dramatization of this possibility in the science fiction movie, *Men in Black*. Will Smith looks at the corpse of a bald giant in the morgue, and touches a little metal latch by the corpse's ear. The face swings open, and there, inside, sitting in a control room, is a little green man. He is the homunculus in the Cartesian theater! We realize now that the corpse is actually a puppet of sorts, being controlled by the little green man, who watches the world on video monitors from the corpse's eyes, and listens to the stereo speakers. I do not think there is anything logically incoherent or self-contradictory about this, and the film carries it off quite well. If we visited a distant planet inhabited by ten-meter-tall intelligent beings, we might want to travel incognito among them by making a ten-meter-tall puppet and driving it around from a control room inside. That would be a Cartesian Theater too. And it could have turned out that when we opened up people's brains we found a little person inside each of them, in the Cartesian Theater pushing all the buttons. But—this is the empirical claim—we do not find that. The conceptual point is that if we did, we would just have to keep peeling off faces until we got to a level that did not have a Cartesian Theater in it, and where the work that would be done by the homunculus is distributed in both space and time within the brain. Once that distribution in space *and time* has been accomplished, Libet's idea can get no purchase. But here is Patrick writing: "Libet produced data that deeply undermined conscious free will." (Haggard and Libet, 2001, p48) Only if you are a Cartesian about conscious free will; otherwise it is just fine.

And here is a particularly vivid case of Cartesians coming out of the closet, "Clearly, conscious intention cannot cause an action if a neural event that precedes and correlates with the act comes before conscious intention." (Roediger, Goode, and Zaromb, 2008, p208)

They think they are just expressing Libet's point of Libet, but let us look more closely. If the neural event is "correlated with" the action, the conscious intention cannot cause the action. This simply does not follow, as can be seen if we change the example. Suppose we are political scientists who have poked around in Washington DC and found that evidence of certain activity by legislators and others was highly predictive of when Congress was going to vote on an issue, and which way Congress was going to vote. We test our theory over many trials and, sure enough, our predictions are borne out more often than not. So we have political events correlated with later voting events. Suppose our Legislative Potential (LP) predicts that Congress will enact a health-care bill after 350 hours, and sure enough, 350 hours later this is just what Congress does. Would that show that the Congressional voting cannot cause the enacting of the law? Of course not, and it is a good thing this would be a mistake, because Soon et. al, in a paper in *Nature Neuroscience* (2008) , found that using more advanced techniques in a Libet-style paradigm, they were able to predict some decisions as much as 10 seconds in advance! When you think about it, this is not all that surprising. It just shows that it takes time to make decisions, and if somebody has the right sort of equipment they can look at those decisions being made and may even be able to predict what they are going to be. Here is one important take-home message from Soon-style experiments: if you are playing poker for big stakes, do not sit in an *f*MRI machine having your brain scanned, because it can reveal your "tells" even before you could say what you are going to do. This does not tell us we do not have free will.

But what *would* give us the free will we want. Consider a useful outburst from Tom Wolfe, a very astute commentator on the American scene: "The conclusion out beyond the laboratory walls is: 'The fix is in. We're all hard-wired' and 'Don't blame me; I'm wired wrong.'" (2001, p100) That is the lesson he gets from neuroscience. "Wired wrong?" Let us take him at his word, and see where it leads. What would it take to be wired right? *Could* we be wired right for responsibility? I think the answer is yes. Good wiring is what provides us with what we could call *moral competence* and it is really quite simple: responsiveness to the representation of reasons and the capacity to recognize and counteract manipulation by other agents (which appear all the time in philosophical thought experiments). Examples of these bogeymen are the puppeteer that controls a body from afar, and the nefarious neurosurgeon who secretly goes into your brain and usurps your autonomy.

Green and Cohen give us a classic example in their essay. This is a thought experiment inspired by the film *The Boys From Brazil*, in which evil Nazis set about making Hitler-clones in Brazil and giving them the same genetics and the same upbringing as *der Fuehrer.* Nasty nefarious neuroscientists indeed, and the thought experiment is intended to reveal the dependence of everyday, common sense thinking on indeterminism.

> Let us suppose then that a group of scientists has managed to create an individual—call him Mr. Puppet—who, by design, engages in some criminal behavior: say a murder done during a drug deal gone bad. (p1780)

Mr. Puppet has been carefully groomed for this role over the years; his education, his daily activities and experiences, and of course his genes, have all been especially tailored by the evil scientists to create the murderer in the dock, who is antisocial, to say the least, but coherent, rational, otherwise well-informed (we are to imagine). Greene and Cohen invite the reader to conclude that Mr Puppet is not responsible for the murder, in spite of his intelligence and apparent competence as a decision-maker. So that is the thought experiment. They say:

> What's the real difference between us and the puppet, Mr. Puppet? One obvious difference is that Mr. Puppet is the victim of a diabolical plot, whereas most people, we presume, are not, but does this matter? . . . . .The thought that Mr. Puppet is not fully responsible that depends on the idea that his actions were externally determined. But the fact that these forces are connected to the desires and intentions of evil scientists is irrelevant, is it not? What matters is only that these forces are beyond Mr. Puppet's control, that they're not really his. (p1780)

That is their telling of it. I was amused to see that they pause in the middle of their thought experiment to say the following in a footnote: "Daniel Dennett might object that the story of Mr. Puppet is just a

misleading intuition pump." Yes, they are right. It is indeed a misleading intuition pump, and I object. Their anticipation of my reaction does not deter them, however. They go on and they say, rather blithely: "It seems to us that the more one knows about Mr. Puppet and his life, the less one is inclined to see him as truly responsible for his actions, and consider our punishing him as a worthy end in itself." Let us take a closer look. As Douglas Hofstadtder has wisely advised, when confronting a new intuition pump, turn all the knobs and see what features are doing all the work. So first, let us get rid of that group of scientists, the manipulators, and replace it with an indifferent environment.

> Let us suppose, then, that *an indifferent environment* has managed to create an individual—call him Mr. Puppet—who, by design, engages in some criminal behavior: say, a murder done during a drug deal gone bad. . . .

Next I want to get rid of "by design" because if we have got an indifferent environment then nobody has designed this individual, so we will replace it with the phrase "with high probability."

> Let us suppose then that an indifferent environment has managed to create an individual—call him Mr. Puppet—who, *with high probability*, engages in some criminal behavior. . . .

Finally, I want to turn one more knob. I want to do something merely cosmetic, entirely trivial: I want to change the fellow's name. Let us see, what shall we call him? How about Captain Autonomy?

> Let us suppose then that an indifferent environment has managed to create an individual—call him Captain Autonomy—who, *with high probability*, engages in some criminal behavior. .

But now we can see that Greene and Cohen are just wrong. The more you consider their intuition pump the more obvious it becomes that it is *not* the imagined determinism, but the secret manipulation by usurping evil agents, that, to our intuitions, destroys or diminishes the poor fellow's responsibility. *If* Captain Autonomy has diminished responsibility, I do not think it has anything to do with determinism or indeterminism, it has to do with some imagined diminished competence, not because his character is externally caused. If his competence is intact, then there is simply no reason to consider his responsibility diminished. We invoke what I call the Principle of Default Responsibility: *if no other agent is responsible for your condition and the acts that flow from it, you are.* The buck stops there, if you are competent.

Then what about moral responsibility and desert? Forward looking consequentialism, as Green and Cohen say, seems to leave this out. Backward looking retributivism seems to require something metaphysically impossible. And Green and Cohen see no way of supporting any just deserts clause that is not retributive and libertarian, that is to say, that does not require indeterminism. But I want to say that there is a consequentialist middle ground. I am going to sketch it out.

Here, then, is a *consequentialist* account of desert. The *consequences* of the policy of holding one's self and others responsible justify the policy. Why? Because the mutual presumptions of competence that this entails support a host of societal activities: trade, promising, multi-agent/multi-year projects. Civilization, in effect, depends on the practice of holding ourselves and others responsible. Those who want to participate in this institution, which is the source of many benefits, must tacitly accept the rules. And what are the rules? First of all, membership is exclusive; you have to be competent, and you have to do what I call "making yourself large." Not only must you make yourself large; you must protect your boundaries from incursions by manipulators. "If you make yourself really small, you can externalize virtually everything." (Dennett, 1983. p143). You *can* externalize everything, but you do not want to; you want to make yourself large and *take* responsibility for your deeds using the default responsibility clause. And then you also want to accept that you are eligible for punishment if you transgress. These are just the norms for responsibility. Call this the *agent club*. Membership is exclusive but open to all who meet the requirements and make the commitments. This agent club is a cousin to Plato's myth of the metals, or the social contract myth, or

John Rawls' 'original position'. I am not saying it is something entirely new. It is a rational reconstruction of the grounds for preserving and defending the just deserts clause.

What is the alternative proposed by Green and Cohen? Consequentialist *medicalization*, in effect, which replaces punishment in all its forms with . . . . treatment and education designed to encourage a revision of one's goals. Greene and Cohen think that this sort of consequentialism is humane, and it is clearly more humane than the horrific vindictiveness of Kant's retributivism, but otherwise I think it is quite horrible in its own right. It is, or generally leads to, totalitarianism. Think of all the people in the Soviet Union who, it was decided by the state, needed "curing" of the mental disorders that caused them to commit anti-state acts. They were not sick. They were enemies of the state; they certainly did not want to be institutionalized for a cure; they actually had the right, according to sane retributivism, to be punished. Not only is the consequentialist alternative to punishment a slippery slope to totalitarianism, but it would tend to erode the trust that our institutions depend on. I am not saying that current practices of punishment are humane – they are in many regards disgusting. We should reform punishment, not abandon it.

When we look at what keeps a system of punishment from eroding into viciousness of either the retributive or consequentialist sorts is what we might call the "arms race of punishment." The law presupposes, not indeterminism, but *moral competence*. Then respect for the law requires that the law make humane *exemptions*; we recognize that some people, under various circumstances, just are not competent, so, responding to the community's disapproval of holding such people fully responsible, the law begins to build in some exemptions. But then those exemptions or exceptions lead to loop-hole-seeking by the craftier of our fellow citizens when they run afoul of the law (exploiting the insanity defense, etc., etc.), and this then leads to further revisions of the law, tightening or closing off the loopholes, and so it goes, back and forth an opponent process in which the law is gradually being refined, with people always looking for ways of getting around the law, and the law having to keep adjusting to keep up with this. Now what neuroscience adds to this process is not any revolution but just new loop-hole candidates. As we learn more and more about people's brains there will no doubt be more grounds for saying "these people, under these specific circumstances, are not properly responsible," and this will quite correctly inspire minimalist revisions in the law, which in turn will stimulate further exploration of the new loopholes, leading to further revisions in the law to prevent it from being exploited and abused. Neuroscience is going to give us new candidates for loopholes, but also new ways of addressing them. This has happened many times in the past. Here is a simple example. We adults are all supposedly sophisticated, so if we do not read the fine print in a contract, that is our fault, a risk we have chosen to take because we are lazy or too trusting. But there are also a lot of unsophisticated people out there that do not know enough to read the fine print, so laws have been made to protect them, in effect, and to punish those who would exploit deceptive contracts. But then, once those laws are in place, the law then puts the responsibility back onto the buyer and it's *caveat emptor* all over again, and back and forth goes the arms race. And the same thing is going to happen to the law.

Here, then, are my conclusions: determinism is a red herring, neuroscience has ominous implications only for closet Cartesians, Mr. Puppet is a defective intuition pump, and there is a consequentialist, compatibilist justification of the just deserts clause. Thank you for your attention.

References

Bloom, Paul, 2006, "My brain made me do it," *Journal of Cognition and Culture*.

Burgess, Anthony, 1962, *A Clockwork Orange.* London UK, William Heinemann

Calvin, William, 1990, *The Ascent of Mind: Ice Age Climates and the Evolution of Intelligence*. New York: Bantam

Crick, Francis, 1994, *The Astonishing Hypothesis*, New York: Scribner's.

Dennett, D. C. 1984, *Elbow room: the varieties of free will worth wanting.* Cambridge, MA: MIT Press.

Dennett, D. C. 1991 *Consciousness explained*. Boston, MA: Little Brown and Co.

Dennett, D. C. 2003 *Freedom evolves*. New York: Viking.

Droege, Paula, (2010) "The Role of Unconsciousness in Free Will" *JConscStudies,* 17, no5-6; *pp58-81*

Fodor, Jerry, 2003, review of Dennett, *Freedom Evolves* (2003), in *London Review of Books*, 5 March, 2003).

Gazzaniga, Michael, 1998, *The Mind's Past*, Berkeley,CA: UC Press.

Green, Joshua, Cohen, Jonathan, (2004) "For the law, neuroscience changes everything and nothing," *Philosophical Transactions of the Royal Society B,* **359**, pp1775-1785

Haggard, Patrick, 2001, in Haggard and B. Libet, "Conscious Intention and Brain Activity" JConsc.Studies vol 8 no 11, nov 2001. pp47-63

Libet, Benjamin, 1985, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action," *Behavioral and Brain Sciences*, Vol. 8, pp. 529-539.

Libet, Benjamin, (1999) in Benjamin Libet, Anthony Freeman, and J. K. B. Sutherland, Editors, *The volitional brain: Towards a neuroscience of free will*. Imprint Academic, 1999.

Moore, Michael S. (2011) "Responsible Choices, Desert-Based Legal Institutions, and the Challenges of Contemporary Neuroscience," ***Social Philosophy and Policy,*** Vol. 29, Issue No. 1
Ramachandran, V. S., (1998) , *New Scientist*, 5 Sep 1998, p. 35

Roediger, Henry, Michael Goode, Franklin Zaromb, "Free Will and the Control of Action," in J. Baer, J. Kaufman, and R. Baumeister, eds., *Are We Free?: Psychology and Free Will* (Oxford: Oxford University Press, 2008), 208.

Siegel, Lee, 1991, *Net of Magic: Wonders and Deceptions in India*, Chicago: Chicago
University Press.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J-D., 2008, "Unconscious determinants of free decisions in the human brain," *Nature Neuroscience*, 11-5

Vohs, K. D., and Schooler, J. 2008, “The value of believing in free will:
Encouraging a belief in determinism increases cheating,” *Psychological Science, 19,* 49-54.

Wegner, Daniel, 2002, *The Illusion of Conscious Will,* Cambridge, MA: MIT Press

Wiggins, David, 1973, "Towards a reasonable libertarianism", in *Essays on Freedom of Action*, ed. Ted Honderich, London, Boston, Rourledge and Kegan Paul.

Wolfe, Tom, 2001, *Hooking Up*. New York: Farrar, Strauss, Giroux.


European
University
Institute
European University Press