SAM HARRIS

THE BLOG

Reflections on FREE WILL

A Review by Daniel C. Dennett

Ethics | Free Will | Neuroscience | Philosophy | Publishing | Self | January 26, 2014

(Photo via Steven Kersting)

Daniel Dennett and I agree about many things, but we do not agree about free will. Dan has been threatening to set me straight on this topic for several years now, and I have always encouraged him to do so, preferably in public and in writing. He has finally produced a review of my book *Free Will* that is nearly as long as the book itself. I am grateful to Dan for taking the time to engage me this fully, and I will respond in the coming weeks.**—SH**

Daniel C. Dennett is the Austin B. Fletcher Professor of Philosophy, and Co-Director of the Center for Cognitive Studies at Tufts University. He is the author of *Breaking the Spell, Freedom Evolves, Darwin's Dangerous Idea, Consciousness Explained*, and many other books. He has received two Guggenheim Fellowships, a Fulbright Fellowship, and a Fellowship at the Center for Advanced Studies in Behavioral Science. He was elected to the American Academy of Arts and Sciences in 1987. His latest book, written with Linda LaScola, *Caught in the Pulpit: Leaving Belief Behind.*

This essay was first published at Naturalism.org and has been crossposted here with permission.

* * *

Sam Harris's *Free Will* (2012) is a remarkable little book, engagingly written and jargon-free, appealing to reason, not authority, and written with passion and moral seriousness. This is not an ivory tower technical inquiry; it is in effect a political tract, designed to persuade us all to abandon what he considers to be a morally pernicious idea: the idea of free will. If you are one of the many who have been brainwashed into believing that you have—or rather, *are*—an (immortal, immaterial) soul who makes all your decisions independently of the causes impinging on your material body and especially your brain, then this is the book for you. Or, if you have dismissed dualism but think that *what you are* is a conscious (but material) *ego*, a witness that inhabits a nook in your brain and chooses, independently of external causation, all your voluntary acts, again, this book is for you. It is a fine "antidote," as Paul Bloom says, to this incoherent and socially malignant illusion. The incoherence of the illusion has been demonstrated time and again in rather technical work by philosophers (in spite of still finding supporters in the profession), but Harris does a fine job of making this apparently unpalatable fact accessible to lay people. Its malignance is due to its fostering the idea of Absolute Responsibility, with its attendant implications of what we might call Guilt-in-the-eyes-of-God for the unfortunate sinners amongst us and, for the fortunate, the arrogant and self-deluded idea of Ultimate Authorship of the good we do. We take too much blame, and too much credit, Harris argues. We, and the rest of the world, would be a lot better off if we took ourselves—our selves—less seriously. We don't have the kind of free will that would ground such Absolute Responsibility for either the harm or the good we cause in our lives.

All this is laudable and right, and vividly presented, and Harris does a particularly good job getting readers to introspect on their own decision-making and notice that it just does not conform to the fantasies of this all too traditional understanding of how we think and act. But some of us have long recognized these points and gone on to adopt more reasonable, more empirically sound, models of decision and thought, and we think we can articulate and defend a more sophisticated model of free will that is not only consistent with neuroscience and introspection but also grounds a (modified, toned-down, non-Absolute) variety of responsibility that justifies both praise and blame, reward and punishment. We don't think this variety of free will is an illusion at all, but rather a robust feature of our psychology and a reliable part of the foundations of morality, law and society. Harris, we think, is throwing out the baby with the bathwater.

He is not alone among scientists in coming to the conclusion that the ancient idea of free will is not just confused but also a

major obstacle to social reform. His brief essay is, however, the most sustained attempt to develop this theme, which can also be found in remarks and essays by such heavyweight scientists as the neuroscientists Wolf Singer and Chris Frith, the psychologists Steven Pinker and Paul Bloom, the physicists Stephen Hawking and Albert Einstein, and the evolutionary biologists Jerry Coyne and (when he's not thinking carefully) Richard Dawkins.

The book is, thus, valuable as a compact and compelling expression of an opinion widely shared by eminent scientists these days. It is also valuable, as I will show, as a veritable museum of mistakes, none of them new and all of them seductive—alluring enough to lull the critical faculties of this host of brilliant thinkers who do not make a profession of thinking about free will. And, to be sure, these mistakes have also been made, sometimes for centuries, by philosophers themselves. But I think we *have* made some progress in philosophy of late, and Harris and others need to do their homework if they want to engage with the best thought on the topic.

I am not being disingenuous when I say this museum of mistakes is valuable; I am grateful to Harris for saying, so boldly and clearly, what less outgoing scientists are *thinking but keeping to themselves*. I have always suspected that many who hold this hard determinist view are making these mistakes, but we mustn't put words in people's mouths, and now Harris has done us a great service by articulating the points explicitly, and the chorus of approval he has received from scientists goes a long way to confirming that they *have* been making these mistakes all along. Wolfgang Pauli's famous dismissal of another physicist's work as "not even wrong" reminds us of the value of crystallizing an ambient cloud of hunches into something that can be *shown* to be wrong. Correcting widespread misunderstanding is usually the work of many hands, and Harris has made a significant contribution.

The first parting of opinion on free will is between *compatibilists* and *incompatibilists*. The latter say (with "common sense" and a tradition going back more than two millennia) that free will is incompatible with *determinism*, the scientific thesis that there are causes for everything that happens. Incompatibilists hold that unless there are "random swerves"[1] that disrupt the iron chains of physical causation, none of our decisions or choices can be truly free. *Being caused* means *not* being free—what could be more obvious? The compatibilists deny this; they have argued, for centuries if not millennia, that once you understand what free will really is (and must be, to sustain our sense of moral responsibility), you will see that free will can live comfortably with determinism—if determinism is what science eventually settles on.

Incompatibilists thus tend to pin their hopes on indeterminism, and hence were much cheered by the emergence of quantum indeterminism in 20th century physics. Perhaps the brain *can* avail itself of undetermined quantum swerves at the sub-atomic level, and thus escape the shackles of physical law! Or perhaps there is some other way our choices could be truly undetermined. Some have gone so far as to posit an otherwise unknown (and almost entirely unanalyzable) phenomenon called *agent causation*, in which free choices are caused somehow by an agent, but not by any event in the agent's history. One exponent of this position, Roderick Chisholm, candidly acknowledged that on this view every free choice is "a little miracle"—which makes it clear enough why this is a school of thought endorsed primarily by deeply religious philosophers and shunned by almost everyone else. Incompatibilists who think we have free will, and therefore determinism must be false, are known as *libertarians* (which has nothing to do with the political view of the same name). Incompatibilists who think that all human choices are determined by prior events in their brains (which were themselves no doubt determined by chains of events arising out of the distant past) conclude from this that we can't have free will, and, hence, are not responsible for our actions.

This concern for varieties of indeterminism is misplaced, argue the compatibilists: free will is a phenomenon that requires

neither determinism nor indeterminism; the solution to the problem of free will lies in realizing this, not banking on the quantum physicists to come through with the right physics—or a miracle. Compatibilism may seem incredible on its face, or desperately contrived, some kind of a trick with words, but not to philosophers. Compatibilism is the reigning view among philosophers (just over 59%, according to the 2009 Philpapers survey) with libertarians coming second with 13% and hard determinists only 12%. It is striking, then, that all the scientists just cited have landed on the position rejected by almost nine out of ten philosophers, but not so surprising when one considers that these scientists hardly ever consider the compatibilist view or the reasons in its favor.

Harris *has* considered compatibilism, at least cursorily, and his opinion of it is breathtakingly dismissive: After acknowledging that it is the prevailing view among philosophers (including his friend Daniel Dennett), he asserts that "More than in any other area of academic philosophy, the result resembles theology." This is a low blow, and worse follows: "From both a moral and a scientific perspective, this seems deliberately obtuse." (18) I would hope that Harris would pause at this point to wonder—just wonder—whether *maybe* his philosophical colleagues had seen some points that had somehow escaped him in his canvassing of compatibilism. As I tell my undergraduate students, whenever they encounter in their required reading a claim or argument that seems just plain stupid, they should probably double check to make sure they are not misreading the "preposterous" passage in question. It is *possible* that they have uncovered a howling error that has somehow gone unnoticed by the profession for generations, but not very likely. In this instance, the chances that Harris has underestimated and misinterpreted compatibilism seem particularly good, since the points he defends later in the book agree right down the line with compatibilism; he himself is a compatibilist in everything but name!

Seriously, his *main* objection to compatibilism, issued several times, is that what compatibilists mean by "free will" is not what everyday folk mean by "free will." Everyday folk mean something demonstrably preposterous, but Harris sees the effort by compatibilists to make the folks' hopeless concept of free will presentable as somehow disingenuous, unmotivated spin-doctoring, not the project of sympathetic reconstruction the compatibilists take themselves to be engaged in. So it all comes down to who gets to decide how to use the term "free will." Harris is a compatibilist about moral responsibility and the importance of the distinction between voluntary and involuntary actions, but he is not a compatibilist about free will since he thinks "free will" has to be given the incoherent sense that emerges from uncritical reflection by everyday folk. He sees quite well that compatibilism is "the only philosophically respectable way to endorse free will" (p. 16) but adds:

> However, the 'free will' that compatibilists defend is not the free will that most people feel they have. (p. 16)

First of all, he doesn't know this. This is a guess, and suitably expressed questionnaires might well prove him wrong. That is an empirical question, and a thoughtful pioneering attempt to answer it suggests that Harris's guess is simply mistaken.[2] The newly emerging field of experimental philosophy (or "X-phi") has a rather unprepossessing track record to date, but these are early days, and some of the work has yielded interesting results that certainly defy complacent assumptions common among philosophers. The study by Nahmias et al. 2005 found substantial majorities (between 60 and 80%) in agreement with propositions that are compatibilist in outlook, not incompatibilist.

Harris's claim that the folk are mostly incompatibilists is thus dubious on its face, and even if it is true, maybe all this shows is that most people are suffering from a sort of illusion that could be replaced by wisdom. After all, most people used to

believe the sun went around the earth. They were wrong, and it took some heavy lifting to convince them of this. Maybe this factoid is a reflection on how much work science and philosophy still have to do to give everyday laypeople a sound concept of free will. We've not yet succeeded in getting them to see the difference between weight and mass, and Einsteinian relativity still eludes most people. When we found out that the sun does not revolve around the earth, we didn't then insist that there is no such thing as the sun (because what the folk mean by "sun" is "that bright thing that goes around the earth"). Now that we understand what sunsets are, we don't call them illusions. They are real phenomena that can mislead the naive.

To see the context in which Harris's criticism plays out, consider a parallel. The folk concept of *mind* is a shambles, for sure: dualistic, scientifically misinformed and replete with miraculous features—even before we get to ESP and psychokinesis and poltergeists. So when social scientists talk about *beliefs* or *desires* and cognitive neuroscientists talk about *attention* and *memory* they are deliberately using cleaned-up, demystified substitutes for the folk concepts. Is this theology, is this deliberately obtuse, countenancing the use of concepts with such disreputable ancestors? I think not, but the case can be made (there are mad dog reductionist neuroscientists and philosophers who insist that minds are illusions, pains are illusions, dreams are illusions, ideas are illusions—all there is is just neurons and glia and the like). The same could be said about *color*, for example. What everyday folk think colors are—if you pushed them beyond their everyday contexts in the paint store and picking out their clothes—is hugely deluded; that doesn't mean that colors are an illusion. They are real in spite of the fact that, for instance, atoms aren't colored.

Here are some more instances of Harris's move:

> We do not have the freedom we think we have. (p. 5)

Who's *we*? Maybe many people, maybe most, think that they have a kind of freedom that they don't and can't have. But that settles nothing. There may be other, better kinds of freedom that people also think they have, and that are worth wanting (Dennett, 1984).

> We do not know what we intend to do until the intention itself arises. [True, but so what?] To understand this is to realize that we are not the authors of our thoughts and actions *in the way that people generally suppose* [my italics]. (p. 13)

Again, so what? Maybe we are authors of our thoughts and actions in a slightly different way. Harris doesn't even consider that possibility (since that would require taking compatibilist "theology" seriously).

> If determinism is true, the future is set—and this includes all our future states of mind and our subsequent behavior. And to the extent that the law of cause and effect is subject to

> indeterminism—quantum or otherwise—we can take no credit for what happens. There is no combination of these truths that seem compatible with *the popular notion of free will* [my italics]. (p. 30)

Again, the *popular* notion of free will is a mess; we knew that long before Harris sat down to write his book. He needs to go after the attempted improvements, and *it cannot be part of his criticism that they are not the popular notion*.

There is also another problem with this paragraph: the sentence about indeterminism is false:

> And to the extent that the law of cause and effect is subject to indeterminism—quantum or otherwise—we can take no credit for what happens.

Here is a counterexample, contrived, but highlighting the way indeterminism could infect our actions and still leave us responsible (a variant of an old—1978—counterexample of mine):

> You must correctly answer three questions to save the world from a space pirate, who provides you with a special answering gadget. It has two buttons marked YES and NO and two foot pedals marked YES and NO. A sign on the gadget lights up after every question "Use the buttons" or "Use the pedals." You are asked "is Chicago the capital of Illinois?", the sign says "Use the buttons" and you press the No button with your finger. Then you are asked "Are Dugongs mammals?", the sign says "Use the buttons" and you press the Yes button with your finger. Finally you are asked "Are proteins made of amino acids?" and the sign says "Use the pedals" so you reach out with your foot and press the Yes pedal. A roar of gratitude goes up from the crowd. You've saved the world, thanks to your knowledge and responsible action! But all three actions were unpredictable by Laplace's demon because whether the light said "Button" or "Pedals" was caused by a quantum random event. In a less obvious way, random perturbations could infect (without negating) your every deed. The tone of your voice when you give your evidence could be tweaked up or done, the pressure of your trigger finger as you pull the trigger could be tweaked greater or lesser, and so forth, without robbing you of responsibility. Brains are, in all likelihood, designed by natural selection to absorb random fluctuations without being seriously diverted by them—just as computers are. But that means that randomness need not destroy the rationality, the well-governedness, the sense-making integrity of your control system. Your brain may even exploit randomness in a variety of ways to enhance its heuristic search for good solutions to problems.

These are not new ideas. For instance I have defended them explicitly in 1978, 1984, and 2003. I wish Harris had noticed that he contradicts them here, and I'm curious to learn how he proposes to counter my arguments.

Another mistake he falls for—in very good company—is the mistake the great J. L. Austin makes in his notorious footnote about his missed putt. First Austin's version, and my analysis of the error, and then Harris's version.

> Consider the case where I miss a very short putt and kick myself because I could have holed it. It is not that I should have holed it if I had tried: I did try, and missed. It is not that I should have holed it if conditions had been different: that might of course be so, but *I am talking about conditions as they precisely were* [my italics], and asserting that I could have holed it. There is the rub. Nor does 'I can hole it this time' mean that I shall hole it this time if I try or if anything else; for I may try and miss, and yet not be convinced that I could not have done it; indeed, *further experiments may confirm my belief that I could have done it that time* [my italics], although I did not. (Austin 1961: 166. ["Ifs and Cans," in Austin, *Philosophical Papers*, edited by J. Urmson and G. Warnock, Oxford, Clarendon Press.])

Austin claims to be talking about conditions as they precisely were, but if so, then further experiments could not confirm his belief. Presumably he has in mind something like this: he could line up ten "identical" putts on the same green and, say, sink nine out of ten. This would show, would it not, that he could have made that putt? Yes, to the satisfaction of almost everybody, but No, if he means under conditions "as they precisely were," for conditions were subtly different in every subsequent putt—the sun a little lower in the sky, the green a little drier or moister, the temperature or wind direction ever so slightly different, Austin himself older and maybe wiser, or maybe more tired, or maybe more relaxed. This variation is not a bug to be eliminated from such experiments, but a feature without which experiments *could not* show that Austin "could have done otherwise," and this is precisely the elbow room we need to see that "could have done otherwise" is perfectly compatible with determinism, because it *never* means, in real life, what philosophers have imagined it means: replay *exactly* the same "tape" and get a different result. Not only can such an experiment never be done; if it could, it wouldn't' show what needed showing: something about Austin's ability as a golfer, which, like all abilities, needs to be demonstrated to be robust under variation.

Here is Harris' version of the same mistake:

> To say that they were free not to rape and murder is to say that they could have resisted the impulse to do so (or could have avoided feeling such an impulse altogether)—with the universe, including their brains, in precisely the same state it was in at the moment they committed their crimes. (p. 17)

Just not true. If we are interested in whether somebody has free will, it is some kind of ability that we want to assess, and you can't assess *any* ability by "replaying the tape." (See my extended argument to this effect in *Freedom Evolves*, 2003) The point was made long ago by A. M. Honoré in his classic paper "Can and Can't," in *Mind*, 1964, and more recently deeply grounded in Judea Pearl's *Causality: Models, Reasoning and Inference*, [CUP] 2000. This is as true of the abilities of automobiles as of people. Suppose I am driving along at 60 MPH and am asked if my car can also go 80 MPH. Yes, I

reply, but not in *precisely* the same conditions; I have to press harder on the accelerator. In fact, I add, it can also go 40 MPH, but not with conditions *precisely* as they are. Replay the tape till eternity, and it will never go 40MPH in just these conditions. So if you want to know whether some rapist/murderer was "free not to rape and murder," don't distract yourself with fantasies about determinism and rewinding the tape; rely on the sorts of observations and tests that everyday folk use to confirm and disconfirm their verdicts about who could have done otherwise and who couldn't.[3]

One of the effects of Harris's misconstruing compatibilism is that when he turns to the task of avoiding the dire conclusions of the hard determinists, he underestimates his task.[4] At the end of the book, he gets briefly concessive, throwing a few scraps to the opposition:

> And it is wise to hold people responsible for their actions when doing so influences their behavior and brings benefit to society. But this does not mean that we must be taken in by the illusion of free will. We need only acknowledge that efforts matter and that people can change. We do not change ourselves, precisely—because we have only ourselves with which to do the changing—but we continually influence, and are influenced by, the world around us and the world within us. It may seem paradoxical to hold people responsible for what happens in their corner of the universe, but once we break the spell of free will, we can do this precisely to the degree that it is useful. Where people can change, we can demand that they do so. Where change is impossible, or unresponsive to demands, we can chart some other course. (p. 63)

Harris should take more seriously the various tensions he sets up in this passage. It is wise to hold people responsible, he says, even though they are not responsible, not *really*. But we don't hold *everybody* responsible; as he notes, we excuse those who are unresponsive to demands, or in whom change is impossible. That's an important difference, and it is based on the different abilities or competences that people have. Some people (are determined to) have the abilities that justify our holding them responsible, and some people (are determined to) lack those abilities. But determinism doesn't do any work here; in particular it doesn't disqualify those we hold responsible from occupying that role. In other words, *real* responsibility, the kind the everyday folk *think* they have (if Harris is right), is strictly impossible; but when those same folk wisely and justifiably *hold* somebody responsible, that isn't real responsibility![5]

And what is Harris saying about whether we can change ourselves? He says we can't change ourselves "precisely" but we can influence (and hence change) others, and they can change us. But then why can't we change ourselves by getting help from others to change us? Why, for that matter, can't we do to ourselves what we do to those others, reminding ourselves, admonishing ourselves, reasoning with ourselves? It does work, not always but enough to make to worth trying. And notice: if we do things to influence and change others, and thereby turn them into something bad—encouraging their racist or violent tendencies, for instance, or inciting them to commit embezzlement, we may be *held* responsible for this socially malign action. (Think of the drunk driving laws that now hold the bartender or the party host partly responsible for the damage done.) But then by the same reasoning we *can* justifiably be held responsible for influencing ourselves, for good or ill. We can take some credit for any improvements we achieve in others—or ourselves—and we can share the blame for any damage we do to others or ourselves.

There are complications with all this, but Harris doesn't even look at the surface of these issues. For instance, our capacities

to influence ourselves are themselves only partly the result of earlier efforts at self-improvement in which we ourselves played a major role. It takes a village to raise a child, as Hilary Clinton has observed. In the end, if we trace back far enough to our infancy or beyond, we arrive at conditions that we were just lucky (or unlucky) to be born with. This undeniable fact is not the disqualifier of responsibility that Harris and others assume. It disqualifies us for "Ultimate" responsibility, which would require us to be—like God!—causa sui, the original cause of ourselves, as Galen Strawson has observed, but this is nonsense. Our lack of Ultimate responsibility is not a moral blemish; if the discovery of this lack motivates some to reform our policies of reward and punishment, that is a good result, but it is hardly compelled by reason.

This emerging idea, that we can justifiably be held to be the authors (if not the Authors) of not only our deeds but the character from which our deeds flow, undercuts much of the rhetoric in Harris's book. Harris is the author of his book; he is responsible for both its virtues, for which he deserves thanks, and its vices, for which he may justifiably be criticized. But then why can we not generalize this point to Harris himself, and rightly hold him at least partly responsible for his character since it too is a product—with help from others, of course—of his earlier efforts? Suppose he replied that he is not *really* the author of *Free Will*. At what point do we get to use Harris's criticism against his own claims? Harris might claim that he is not really responsible, isn't really the author of his own book, isn't really responsible, *but that isn't what the folk would say. The folk believe in a kind of responsibility that is exemplified by Harris's authorship*. Harris would have distorted the folk notion of responsibility as much if not more than compatibilists have distorted the folk notion of free will.

Harris opens his book with an example of murderous psychopaths, Hayes and Komisarjevsky, who commit unspeakable atrocities. One has shown remorse, the other reports having been abused as a child.

> Whatever their conscious motives, these men cannot know why they are as they are. Nor can we account for why we are not like them.

Really? I think we can. The sentence is ambiguous, in fact. Harris knows full well that we can provide detailed and empirically supported accounts of why normal, law-abiding people who would never commit those atrocities emerge by the millions from all sorts of backgrounds, and why these psychopaths are different. But he has a different question in mind: why we—you and I—are in the fortunate, normal class instead of having been doomed to psychopathy. A different issue, but also an irrelevant, merely *metaphysical* issue. (Cf. "Why was I born in the 20th century, and not during the Renaissance? We'll never know!")

The rhetorical move here is well-known, but indefensible. If you're going to raise these horrific cases, it behooves you to consider that they might be cases of pathology, as measured against (moral) health. Lumping the morally competent with the morally incompetent and then saying "there *really* is no difference between them, is there?" is a move that needs support, not something that can be done by assumption or innuendo.

> I cannot take credit for the fact that I don't have the soul of a psychopath. (p. 4)

True—and false. Harris can't take credit for the luck of his birth, his having had a normal moral education—that's just

luck—but those born thus lucky are informed that they have a duty or obligation to preserve their competence, and grow it, and educate themselves, and Harris has responded admirably to those incentives. He *can* take credit, not Ultimate credit, whatever that might be, but partial credit, for husbanding the resources he was endowed with. As he says, he is just lucky not to have been born with Komisarjevsky's genes and life experiences. If he had been, he'd have been Komisarjevsky!

A similar difficulty infects his claim that there is no difference between an act caused by a brain tumor and an act caused by a belief (which is just another brain state, after all).

> But a neurological disorder appears to be just a special case of physical events giving rise to thoughts and actions. Understanding the neurophysiology of the brain, therefore, would seem to be as exculpatory as finding a tumor in it. (p. 5)

Notice the use of "appears" and "seem" here. Replace them both with "is" and ask if he's made the case. (In addition to the "surely"-alarm I recommend all readers install in their brains (2013), a "seems"–alarm will pick up lots of these slippery places where philosophers defer argument where argument is called for.

Even the simplest and most straightforward of Harris's examples wilt under careful scrutiny:

> Did I consciously choose coffee over tea? No. The choice was made for me by events in my brain that I, as the conscious witness of my thoughts and actions, could not inspect or influence. (p. 7)

Not so. He can influence those internal, unconscious actions—by reminding himself, etc. He just can't influence them *at the moment they are having their effect on his choice*. (He also can't influence the unconscious machinery that determines whether he returns a tennis serve with a lob or a hard backhand once the serve is on its way, but that doesn't mean his tennis strokes are involuntary or outside his—indirect—control. At one point he says "If you don't know what your soul is going to do, you are not in control." (p. 12) Really? When you drive a car, are you not in control? You know "your soul" is going to do the right thing, whatever in the instant it turns out to be, and that suffices to demonstrate to you, and the rest of us, that you are in control. Control doesn't get any more real than that.)

Harris ignores the reflexive, repetitive nature of thinking. My choice at time t can influence my choice at time t' which can influence my choice at time t". How? My choice at t can have among its effects the biasing of settings in my brain (which I cannot directly inspect) that determine (I use the term deliberately) my choice at t'. I *can* influence my choice at t'. I influenced it at time t (without "inspecting" it). Like many before him, Harris shrinks the *me* to a dimensionless point, "the witness" who is stuck in the Cartesian Theater awaiting the decisions made elsewhere. That is simply a bad theory of consciousness.

> I, as the conscious witness of my experience, no more initiate events in my prefrontal cortex than I

> cause my heart to beat. (p. 9)

If this isn't pure Cartesianism, I don't know what it is. His prefrontal cortex is *part of* the I in question. Notice that if we replace the "conscious witness" with "my brain" we turn an apparent truth into an obvious falsehood: "My brain can no more initiate events in my prefrontal cortex than it can cause my heart to beat."

There are more passages that exhibit this curious tactic of heaping scorn on daft doctrines of his own devising while ignoring reasonable compatibilist versions of the same ideas, but I've given enough illustrations, and the rest are readily identifiable once you see the pattern. Harris clearly thinks compatibilism is not worth his attention (so "deliberately obtuse" is it), but after such an indictment, he better come up with some impressive criticisms. His main case against compatibilism—aside from the points above that I have already criticized—consists of three rhetorical questions lined up in a row (pp. 18-19). Each one collapses on closer inspection. As I point out in *Intuition Pumps and Other Tools for Thinking*, rhetorical questions, which are stand-ins for *reductio ad absurdum* arguments so obvious that they need not be spelled out, should always be scrutinized as likely weak spots in arguments.. I offer Harris's trio as exhibits A,B, and C:

> (A) You want to finish your work, but you are also inclined to stop working so that you can play with your kids. You aspire to quite smoking, but you also crave another cigarette. You are struggling to save money, but you are also tempted to buy a new computer. Where is the freedom when one of these opposing desires *inexplicably* [my italics] triumphs over its rival?

But no compatibilist has claimed (so far as I know) that our free will is absolute and trouble-free. On the contrary there is a sizable and fascinating literature on the importance of the various well-known ways in which we respond to such looming cases of "weakness of will," from which we all suffer. When one desire triumphs, this is not usually utterly inexplicable, but rather the confirmable result of efforts of self-manipulation and self-education, *based on empirical self-exploration*. We learn something about what makes us tick—not usually in neuroscientific terms, but rather in terms of folk psychology—and design a strategy to correct the blind spots we find, the biases we identify. That practice undeniably occurs, and undeniably works to a certain extent. We *can* improve our self-control, and this is a morally significant fact about the competence of normal adults—the only people whom we hold fully (but not "absolutely" or "deeply") responsible. Remove the word "inexplicably" from exhibit A and the rhetorical question has a perfectly good answer: in many cases our freedom is an achievement, for which we are partly responsible. (Yes, luck plays a role but so does skill; we are not *just* lucky. (Dennett, 1984)

> (B) The problem for compatibiism runs deeper, however—for where is the freedom in wanting what one wants without any internal conflict whatsoever?

To answer a rhetorical question with another, so long as one can get what one wants so wholeheartedly, what could be better? What could be more freedom than that? Any realistic, reasonable account of free will acknowledges that we are

stuck with some of our desires: for food and comfort and love and absence of pain—and the freedom to do what we want. We can't not want these, or if we somehow succeed in getting ourselves into such a sorry state, we are pathological. These are the healthy, normal, sound, wise desires on which all others must rest. So banish the fantasy of any account of free will that is screwed so tight it demands that we aren't free unless *all* our desires and meta-desires and meta-meta-desires are optional, choosable. Such "perfect" freedom is, of course, an incoherent idea, and if Harris is arguing against it, he is not finding a "deep" problem with compatibilism but a shallow problem with his incompatibilist vision of free will; he has taken on a straw man, and the straw man is beating him.

> (C) Where is the freedom in being perfectly satisfied with your thoughts, intentions, and subsequent actions when they are the product of prior events that you had absolutely no hand in creating?

Not only has he not shown that you had absolutely no hand in creating those prior events, but it is false, as just noted. Once you stop thinking of free will as a magical metaphysical endowment and start thinking of it as an explicable achievement that individual human beings normally accomplish (very much aided by the societies in which they live), much as they learn to speak and read and write, this rhetorical question falls flat. Infants don't have free will; normal adults do. Yes, those of us who have free will are lucky to have free will (we're lucky to be human beings, we're lucky to be alive), but our free will is not just a given; it is something we *are obliged* to protect and nurture, with help from our families and friends and the societies in which we live.

Harris allows himself one more rhetorical question on page 19, and this one he emphatically answers:

> (D) Am I free to do *that which does not occur to me to do?* Of course not.

Again, really? You're playing bridge and trying to decide whether or not to win the trick in front of you. You decide to play your ace, winning the trick. Were you free to play a low card instead? *It didn't occur to you* (it should have, but you acted rather thoughtlessly, as your partner soon informs you). Were you free to play your six instead? In some sense. We wouldn't play games if there weren't opportunities in them to make one choice or another. But, comes the familiar rejoinder, if determinism is true and we rewound the tape of time and put you in exactly the same physical state, you'd ignore the six of clubs again. True, but so what? It does not show that you are not the agent you think you are. *Contrast* your competence at this moment with the "competence" of a robotic bridge-playing doll that *always* plays its highest card in the suit, no matter what the circumstances. It wasn't free to choose the six, because it would play the ace *whatever the circumstances were* whereas if it occurred to you to play the six, you could do it, depending on the circumstances. Freedom involves the ability to have one's choices influenced by changes in the world that matter under the circumstances. Not a perfect ability, but a reliable ability. If you are such a terrible bridge player that you can never see the virtue in ducking a trick, playing less than the highest card in your hand, then your free will at the bridge table is seriously abridged: you are missing the opportunities that make bridge an interesting game. If determinism is true, are these real opportunities? Yes, as real as an opportunity could be: thanks to your perceptual apparatus, your memory, and the well-lit environment, you are

caused/determined to evaluate the situation as one that calls for playing the six, and you play the six.

Turn to page 20 and get one more rhetorical question:

> (E) And there is no way I can influence my desires—for what tools of influence would I use? Other desires?

Yes, for starters. Once again, Harris is ignoring a large and distinguished literature that defends this claim. We use the same tools to influence our own desires as we use to influence other people's desires. I doubt that he denying that we ever influence other people's desires. His book is apparently an attempt to influence the beliefs and desires of his readers, and it seems to have worked rather better than I would like. His book also seems to have influenced his own beliefs and desires: writing it has blinded him to alternatives that he really ought to have considered. So his obliviousness is something for which he himself is partly responsible, having labored to create a mindset that sees compatibilism as deliberately obtuse.

When Harris turns to a consideration of my brand of compatibilism, he quotes at length from a nice summary of it by Tom Clark, notes that I have approved of that summary, and then says that it perfectly articulates the difference between my view and his own. And this is his rebuttal:

> As I have said, I think compatibilists like Dennett change the subject: They trade a psychological fact—the subjective experience of being a conscious agent—for a conceptual understanding of ourselves as persons. This is a bait and switch. The psychological truth is that people feel identical to a certain channel of information in their conscious minds. Dennett is simply asserting that we are more than this—we are coterminous with everything that goes on inside our bodies, whether we are conscious of it or not. This is like saying we are made of stardust—which we are. But we don't feel like stardust. And the knowledge that we are stardust is not driving our moral intuitions or our system of criminal justice. (p. 23)

I have thought long and hard about this passage, and I am still not sure I understand it, since it seems to be at war with itself. Harris apparently thinks you see yourself as a conscious witness, perhaps immaterial—an immortal soul, perhaps—that is distinct from (the rest of?) your brain. He seems to be saying that this folk understanding people have of *what they are identical to* must be taken as a "psychological fact" that anchors any discussion of free will. And then he notes that I claim that this folk understanding is just plain wrong and try to replace it with a more scientifically sound version of what a conscious person is. Why is it "bait and switch" if I claim to *improve* on the folk version of personhood before showing how it allows for free will? He can't have it both ways. He is certainly claiming in his book that the dualism that is uncritically endorsed by many, maybe most, people is incoherent, and he is right—I've argued the same for decades. But then how can he object that I want to replace the folk conception of free will based on that nonsense with a better one? The fact that the folk don't *feel* as if they are larger than their imagined Cartesian souls doesn't count against my account, since I am proposing to correct the mistake manifest in that "psychological fact" (if it is one). And if Harris thinks that it is this folk

notion of free will that “drives our moral intuitions and our legal system” he should tackle the large literature that says otherwise. (starting with, e.g., Stephen Morse[6]).

One more rhetorical question:

> (G) How can we be ‘free’ as conscious agents if everything that we consciously intend is caused by events in our brain that we *do not* intend and of which we are entirely unaware? We can’t. (p. 25-26)

Let’s take this apart, separating its elements. First let’s try dropping the last clause: “of which we are entirely unaware”.

> How can we be ‘free’ as conscious agents if everything that we consciously intend is caused by events in our brain that we do not intend?

Well, if the events that cause your intentions are thoughts about what the best course of action probably is, and why it is the right thing to do, then that causation strikes me as the very epitome of freedom: you have the ability to intend exactly what you think to be the best course of action. When folks lack that ability, when they find they are unable to act intentionally on the courses of action they deem best, all things considered, we say they suffer from weakness of will. An intention that was an apparently causeless orphan, arising for no discernible reason, would hardly be seen as free; it would be viewed as a horrible interloper, as in alien hand syndrome, imposed on the agent from who knows where.

Now let’s examine the other half of Harris’s question:

> How can we be “free” as conscious agents if everything that we consciously intend is caused by events in our brain of which we are entirely unaware?

I don’t always have to reflect, consciously, on my reasons for *my* intentions for them to be both mine and free. When I say “thank you” to somebody who gives me something, it is “force of habit” and I am entirely unaware of the events in my brain that cause me to say it but it is nonetheless a good example of a free action. Had I had a reason to override the habit, I would have overridden it. My not doing so tacitly endorses it as an action of mine. Most of the intentions we frame are like this, to one degree or another: we “instinctively” reach out and pull the pedestrian to safety without time for thinking; we rashly adopt a sarcastic tone when replying to the police officer, we hear the doorbell and jump up to see who’s there. These are all voluntary actions for which we are normally held responsible if anything hinges on them. Harris notes that the voluntary/involuntary distinction is a valuable one, but doesn’t consider that it might be part of the foundation of our moral and legal understanding of free will. Why not? Because he is so intent on bashing a caricature doctrine.

He ends his chapter on compatibilism with this:

> People *feel* that they are the authors of their thoughts and actions, and this is the only reason why there seems to be a problem of free will worth talking about. (p. 26)

I can agree with this, if I am allowed to make a small insertion:

> People *feel* that they are the authors of their thoughts and actions, *and interpreted uncharitably, their view can be made to appear absurd; taken the best way, however, they can be right;* and this is the only reason why there seems to be a problem of free will worth talking about.

One more puzzling assertion:

> Thoughts like "What should I get my daughter for her birthday? I know—I'll take her to a pet store and have her pick out some tropical fish" convey the apparent reality of choices, freely made. But from a deeper perspective (speaking both objectively and subjectively) thoughts simply arise unauthored and yet author our actions. (p. 53)

What would an authored thought look like, pray tell? And how can unauthored thoughts author our actions? Does Harris mean *cause, shape and control* our actions? But if an unauthored thought can cause, shape and control something, why can't a whole person cause, shape and control something? Probably this was misspeaking on Harris's part. He should have said that unauthored thoughts are the causes, shapers and controllers—but not the authors—of our actions. Nothing could be an author, not really. But here again Harris is taking an everyday, folk notion of authorship and inflating it into metaphysical nonsense. If he can be the author of his book, then he can be the author of his thoughts. If he is not the author of *Free Will*, he should take his name off the cover, shouldn't he? But he goes on immediately to say he *is* the cause of his book, and "If I had not decided to write this book, it wouldn't have written itself."

> Decisions, intentions, efforts, goals, willpower, etc., are causal states of the brain, leading to specific behaviors, and behaviors lead to outcomes in the world. Human choice, therefore, is as important as fanciers of free will believe. But the next choice you make will come out of the darkness of prior causes that you, the conscious witness of your experience, did not bring into being. (p. 34)

We've already seen that the last sentence is false. But notice that *if it were true*, then it would be hard to see why "human

choice is important"—except in the way lightning bolts are important (they can do a lot of damage). If your choices "come out of the darkness" and you did not bring them into being, then they are like the involuntary effusions of sufferers from Tourette's Syndrome, who blurt out obscenities and make gestures that are as baffling to them as to others. In fact we know very well that I can influence your choices, and you can influence my choices, and even your own choices, and that this "bringing into being" of different choices is what makes them morally important. That's why we exhort and chastise and instruct and praise and encourage and inform others and ourselves.

Harris draws our attention to how hard it can be to change our bad habits, in spite of reading self-help books and many self-admonitions. These experiences, he notes, "are not even slightly suggestive of freedom of the will" (p. 35). True, but then other experiences we have are often very suggestive of free will. I make a promise, I solemnly resolve to keep it, and happily, I do! I hate grading essays, but recognizing that my grades are due tomorrow, I reluctantly sit down and grind through them. I decide to drive to Boston and lo and behold, the next thing I know I'm behind the wheel of my car driving to Boston! If I could almost never do such things I would indeed doubt my own free will, and toy with the sad conclusion that somewhere along the way I had become a helpless victim of my lazy habits and no longer had free will. Entirely missing from Harris's account—and it is not a lacuna that can be repaired—is any acknowledgment of the morally important difference between normal people (like you and me and Harris, in all likelihood) and people with serious deficiencies in self-control. The reason he can't include this missing element is that his whole case depends in the end on insisting that there really is no morally relevant difference between the raving psychopath and us. We have no more *free will* than he does. Well, we have more *something* than he does, and it is morally important. And it looks very much like what everyday folks often call free will.

> Of course you can create a framework in which certain decisions are more likely than others—you can, for instance, purge your house of all sweets, making it very unlikely that you will eat dessert later in the evening—but you cannot know why you were able to submit to such a framework today when you weren't yesterday. (p. 38)

Here he seems at first to be acknowledging the very thing I said was missing in his account above—the fact that you can take steps to bring about an alteration in your circumstances that makes a difference to your subsequent choices. But notice that his concession is short-lived, because he insists that you are just as in the dark about how your decision to purge your house of all sweets came about. But that is, or may well be, false. You may know exactly what train of thought led you to that policy. *But then, you can't know why that train of thought occurred to you, and moved you then*. No, you can, and often do. Maybe your candy-banishing is the *nth*level result of your deciding to decide to decide to decide to decide . . . . to do something about your health. *But since the regress is infinite, you can't be responsible!* Nonsense. You can't be "*ultimately* responsible" (as Galen Strawson has argued) but so what? You can be partially, largely responsible.

I cannot resist ending this catalogue of mistakes with the one that I find most glaring: the cover of Harris's little book, which shows marionette strings hanging down. The point, which he reiterates several times in the book, is that the prior causes (going back to the Big Bang, if you like) that determine your choices *are like* the puppeteer who determines the puppet's every action, every "decision." This analogy enables him to get off a zinger:

> Compatibilism amounts to nothing more than an assertion of the following creed: *A puppet is free as long as he loves his strings.* (p. 20)

This is in no way supported by anything in his discussion of compatibilism. Somehow Harris has missed one of the deepest points made by Von Neumann and Morgenstern in their introduction to their ground-breaking 1953 book, *Theory of Games and Economic Behavior*, [Princeton UP, John and Oskar]. Whereas Robinson Crusoe alone on his desert island can get by with probabilities and expected utility theory, as soon as there is a second agent to deal with, he needs to worry about feedback, secrecy and the intentions of the other agent or agents (what I have called *intentional systems*). For this he needs game theory. There is a fundamental difference between an environment with no competing agents and an environment populated with would-be manipulators.[7] The manifold of causes that determine our choices only intermittently includes other agents, and when they are around they do indeed represent a challenge to our free will, since they may well try to read our minds and covertly influence our beliefs, but the environment *in general* is not such an agent, and hence is no puppeteer. When sunlight bouncing off a ripe apple causes me to decide to reach up and pick it off the tree, I am not being *controlled* by that master puppeteer, Captain Worldaroundme. I am controlling myself, thanks to the information I garner from the world around me. Please, Sam, don't feed the bugbears. (Dennett, 1984)

Harris half recognizes this when later in the book he raises puppets one more time:

> It is one thing to bicker with your wife because you are in a bad mood; it is another to realize that your mood and behavior have been caused by low blood sugar. This understanding reveals you to be a biochemical puppet, of course, but it also allows you to grab hold of one of your strings. A bite of food may be all that your personality requires. Getting behind our conscious thoughts and feelings can allow us to steer a more intelligent course through our lives (while knowing, of course, that we are ultimately being steered). (p. 47)

So unlike the grumpy child (or moody bear), we intelligent human adults can "grab hold of one of our strings". But then if our bodies are the puppets and we are the puppeteers, we *can* control our bodies, and thereby our choices, and hence can be held responsible—really but not Ultimately responsible—for our actions and our characters. We are not immaterial souls but embodied rational agents, determined (in two senses) to do what is right, most of the time, and ready to be held responsible for our deeds.

Harris, like the other scientists who have recently mounted a campaign to convince the world that free will is an illusion, has a laudable motive: to launder the ancient stain of Sin and Guilt out of our culture, and abolish the cruel and all too usual punishments that we zestfully mete out to the Guilty. As they point out, our zealous search for "justice" is often little more than our instinctual yearning for retaliation dressed up to look respectable. The result, especially in the United States, is a barbaric system of imprisonment—to say nothing of capital punishment—that should make all citizens ashamed. By all means, let's join hands and reform the legal system, reduce its excesses and restore a measure of dignity—and freedom!—to those whom the state must punish. But the idea that all punishment is, in the end, unjustifiable and should be *abolished* because nobody is ever *really* responsible, because nobody has "real" free will is not only not supported by science or

philosophical argument; it is blind to the chilling lessons of the not so distant past. Do we want to medicalize all violators of the laws, giving them indefinitely large amounts of involuntary "therapy" in "asylums" (the poor dears, they aren't responsible, but for the good of the society we have to institutionalize them)? I hope not. But then we need to recognize the powerful (consequentialist) [8] arguments for maintaining a system of punishment (and reward). Punishment can be fair, punishment can be justified, and in fact, our societies could not manage without it.

This discussion of punishment versus medicalization may seem irrelevant to Harris's book, and an unfair criticism, since he himself barely alludes to it, and offers no analysis of its possible justification, but that is a problem for him. He blandly concedes we will—and should—go on holding some people responsible but then neglects to say what that involves. Punishment and reward? If not, what does he mean? If so, how does he propose to regulate and justify it? I submit that if he had attempted to address these questions he would have ended up with something like this:

> Those eligible for punishment and reward are those with the *general abilities* to respond to reasons (warnings, threats, promises) rationally. Real differences in these abilities are empirically discernible, explicable, and morally relevant. Such abilities can arise and persist in a deterministic world, and they are the basis for a justifiable policy of reward and punishment, which brings society many benefits—indeed makes society possible. (Those who lack one or another of the abilities that constitute this moral competence are often said, by everyday folk, to lack free will, and this fact is the heart for compatibilism.)

If you think that the fact that *incompatibilist* free will is an illusion demonstrates that no punishment can ever be truly deserved, think again. It may help to consider all these issues in the context of a simpler phenomenon: sports. In basketball there is the distinction between ordinary fouls and flagrant fouls, and in soccer there is the distinction between yellow cards and red cards, to list just two examples. Are these distinctions fair? Justified? Should Harris be encouraged to argue that there is no real difference between the dirty player and the rest (and besides, the dirty player isn't responsible for being a dirty player; just look at his upbringing!)? Everybody who plays games must recognize that games without strictly enforced rules are not worth playing, and the rules that work best do not make allowances for differences in heritage, training, or innate skill. So it is in society generally: we are all considered equal under the law, presumed to be responsible until and unless we prove to have some definite defect or infirmity that robs us of our free will, as ordinarily understood.

Notes

1. The random swerve or *clinamen* is an idea going back to Lucretius more than two thousand years ago, and has been seductive ever since.

- Eddy Nahmias , Stephen Morris , Thomas Nadelhoffer & Jason Turner, 2005, "Surveying Freedom: Folk Intuitions about free will and moral responsibility," *Philosophical Psychology*, 18, pp 561-584
- Given the ocean of evidence that people assess human abilities, including their abilities to do or choose otherwise, by methods that make no attempt to clamp conditions "precisely as they were," overlooking this prospect has required nearly superhuman self-blinkering by incompatibilists. I consider Austin's mistake to be the central core of the ongoing confusion about free will; if you look at the large and intricate philosophical literature about incompatibilism, you will see that just about everyone assumes, without argument, that it is *not* a mistake. Without that assumption the interminable discussions of van Inwagen's "Consequence Argument" could not be formulated, for instance. The excellent article on "Arguments for Incompatibilism" in the online Stanford

Encyclopedia of Philosophy, cites Austin's essay but does not discuss this question.

- Here more than anywhere else we can be grateful to Harris for his forthrightness, since the distinguished scientists who declare that free will is an illusion almost never have much if anything to say about how they think people should treat each other in the wake of their discovery. If they did, they would land in the difficulties Harris encounters. If nobody is responsible, not really, then not only should the prisons be emptied, but no contract is valid, mortgages should be abolished, and we can never hold anybody to account for anything they do. Preserving "law and order" without a concept of real responsibility is a daunting task. Harris at least recognizes his—dare I say?—responsibility to deal with this challenge.
- "I'm writing a book on magic," I explain, and I'm asked, "Real magic?" By *real magic* people mean miracles, thaumaturgical, 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*. (p. 425) – Lee Siegel, *Net of Magic*
- Morse, "The Non-Problem of Free Will in Forensic Psychiatry and Psychology," *Behavioral Sciences and the Law*, Vol. 25 (2007), pp. 203-220; Morse, "Determinism and the Death of Folk Psychology: Two Challenges to Responsibility from Neuroscience," *Minnesota Journal of Law, Science, and Technology*, Vol. 9 (2008), pp. 1-36, at pp. 3-13.
- 2.2.2. Crusoe is given certain physical data (wants and commodities) and his task is to combine and apply them in such a fashion as to obtain a maximum resulting satisfaction. There can be no doubt that he controls exclusively all the variables upon which this result depends—say the allotting of resources, the determination of the uses of the same commodity for different wants, etc. Thus Crusoe faces an ordinary maximum problem, the difficulties of which are of a purely technical—and not conceptual—nature, as pointed out.

  2.2.3. Consider now a participant in a social exchange economy. His problem has, of course, many elements in common with a maximum problem. But it also contains some, very essential, elements of an entirely different nature. He too tries to obtain an optimum result. But in order to achieve this, he must enter into relations of exchange with others. If two or more persons exchange goods with each other, then the result for each one will depend in general not merely upon his own actions but on those of the others as well. Thus each participant attempts to maximize a function (his above-mentioned "result") of which he does not control all variables. This is certainly no maximum problem, but a peculiar and disconcerting mixture of several different maximum problems. Every participant is guided by another principle and neither determines all variables which affect his interest.

  This kind of problem is nowhere dealt with in classical mathematics. (Von Neumann and Morgenstern, pp10-11)
- Apparently some thinkers have the idea that any justification of *punishment* is (by definition?) *retributive*. But this is a mistake; there are consequentialist justifications of the “retributive” ideas of *just deserts* and the *mens rea* requirement for guilt, for instance. Consider how one can defend the existence of the red card/yellow card distinction in soccer on purely consequentialist grounds.
  Find this article online at: https://www.samharris.org/blog/item/reflections-on-free-will