Mark Balaguer
Department of Philosophy
Cal State LA

Sapolsky-Freedom and Libertarian-Freedom

1. SAPOLSKY-FREEDOM

Consider the following kind of freedom:

> A person is *Sapolsky-free* just in case at least some of their decisions have the following trait: which option was selected in the decision wasn't causally influenced by any events in the history of the universe.

Robert Sapolsky (2023) argues convincingly for the claim that *we aren't Sapolsky-free*. Sapolsky *claims* to have argued that we don't have *free will*, but he hasn't; that would require him to also argue for the conceptual-analysis claim that *free will is Sapolsky-freedom*—and he hasn't done that.[1]

Also, it's obvious that free will *isn't* Sapolsky-freedom. If it were, then the following sentence would be a contradiction (and it clearly isn't): 'My decision was influenced by some things I learned from my mother, but I chose of my own free will.'

So Sapolsky hasn't argued that we don't have free will; he's just argued that we don't have Sapolksy-freedom. And, for the record, I'm not complaining about this. My own view is we shouldn't bog down in the conceptual-analysis question—i.e., the question of what free will *is*—because that question is just about *words*. More precisely, I think the answer to that question is completely determined by facts about what ordinary folk mean by 'free will'—and I don't think anything important turns on this. I think the right methodology to use here (if we're interested in the nature of human decision-making processes and not in words) is the following:

> *Step 1*: Introduce (and stipulatively define) new terms of art that denote the specific kinds of freedom that we want to talk about (e.g., 'Sapolsky-freedom', 'Hume-freedom', 'Frankfurt-freedom', etc.). *Step 2*: Try to determine which of these kinds of freedom we possess.

[1] Sapolsky doesn't use 'Sapolsky-free'; he uses 'free will'; moreover, he doesn't define that expression—he just says a few hand-wavy things about what he means. I have back-engineered this definition from his remarks.

So when Sapolsky stipulatively defines a kind of freedom and then argues that we don't have that kind of freedom, I think he's using the right methodology (although, again, I think he should've used a term of art like 'Sapolsky-freedom' instead of 'free will').

In addition to using the right methodology, I think Sapolsky's argument against the claim that we're Sapolsky-free is a *good* argument—I think he's *right*.

The problem, though, is that nobody worth listening to has ever believed that we're Sapolsky-free. So Sapolsky's result isn't very interesting. Of *course* we're not Sapolsky-free. Who ever thought otherwise?

More importantly, I don't think that Sapolsky-freedom is *desirable*, or *worth wanting*. Why would I want to be able to make decisions that aren't influenced by my own preferences and deliberations, or by the many things I've learned in life from my parents and teachers and friends? I wouldn't. And I don't.

It doesn't follow from these remarks, however, that Sapolsky's arguments aren't important. For his arguments might give us good reason to doubt that we have *other* kinds of freedom—kinds that are more interesting and/or desirable than Sapolsky-freedom.

Now, it's obvious that Sapolsky's argument don't give us any reason to doubt that we have various *compatibilist* kinds of freedom—e.g., *Hume-freedom*—where a person is *Hume-free* just in case they have the ability to *do what they want*, or to *act on their desires*. It's completely obvious that we're Hume-free, and nothing in Sapolsky's book gives us any reason to doubt this.[2] (And it's worth noting that Hume-freedom is definitely *worth wanting*; indeed, life without Hume-freedom would be a veritable *hell*.)

But the claim that I want to argue for here is that Sapolsky's arguments don't even give us reason to doubt that we have *libertarian* freedom.

2. L-FREEDOM

Consider the following incompatibilist/libertarian kind of freedom:

> A person P is *L-free* just in case P makes at least some decisions that are such that (i) they're not determined (where an event E is determined just in case its occurrence was

[2] Sapolsky argues at length that our intentions and desires and so on are determined by prior events; but this is perfectly compatible with the thesis that we're Hume-free because all that thesis requires is that we're able to do what we want—it doesn't say anything about where our wants come from.

causally necessitated by prior events together the laws of nature); and (ii) they're appropriately non-random (where a decision D counts as *appropriately non-random* just in case it was made *by the agent* in some relevant sense—so that, e.g., the agent was the *source* of D, or the agent *controlled which option was selected* in D, or some such thing); and (iii) the indeterminacy is relevant to the appropriate non-randomness in the sense that it *generates* the non-randomness, or *procures* it, or *enhances* it, or some such thing. (More simply (but less precisely): P is *L-free* just in case some of their decisions are such that (a) *P did it*—i.e., it was *P* who made the decision and not something else—and (b) nothing made P choose in the way that they did, and (c) it's at least partly because of (b) that (a) is true.)

I don't think Sapolsky's arguments give us any reason to doubt that we're L-free. I'll argue for this by articulating a specific version of the view that we're L-free—which I'll call *torn-D-libertarianism*—and then arguing that Sapolsky's arguments don't undermine that view at all. (To be clear, I don't think we have any good reason to think that we *are* L-free; but I also don't think we have any good reason to *doubt* that we're L-free, and, in particular, I don't think Sapolsky has given us such a reason.)

3. TORN-D-LIBERTARIANISM

3.1 The (relatively) uncontroversial part: I'll begin by articulating two theses that are built into the torn-D-libertarian view and that are, I think, unproblematic from a determinist point of view. The two theses are as follows:

(i) Mind-brain materialism is true. More specifically, token-identity theory is true, and in particular, every conscious decision *is* a neural event.

(ii) If we have L-freedom at all, then we have it in connection with (at least some of our) *conscious decisions*—in particular, (at least some of our) *torn decisions*—where a *torn decision* is a decision in which (a) the agent feels completely torn between two or more live options—i.e., the agent has two or more tied-for-best options, and they have no conscious belief about which of them is best (they seem equally good)—and (b) the agent decides *while feeling torn* (presumably because *just choosing* is better for some reason than remaining undecided).

To see why torn-D-libertarians endorse thesis (ii), consider the following two definitions:

> A *non-decisional action* is an action that's not preceded by a decision. (E.g., if you're strolling through a park, daydreaming, you don't consciously decide to move your legs on every step.)
>
> A *no-brainer decision* is a decision in which it seems to the agent, in their conscious thinking, that there's a *unique best option*, and the agent feels *certain* that they should choose that option.

Torn-D-libertarians think that non-decisional actions and no-brainer decisions are either determined or close to determined[3]—and so they don't think we exercise L-freedom in these kinds of cases—and this is why they think the question of whether we're L-free boils down to a question about our torn decisions.[4]

It's important to note that torn decisions aren't defined as being undetermined, and so determinists shouldn't have any objection to the claim that we *make* decisions of this kind. Now, I don't think we make a *lot* of torn decisions, but I do think we make a few of these decisions every day. E.g., you might make a torn decision about whether to go for a beer with a friend or continue working on a paper, or about whether to order chocolate or vanilla at an ice cream parlor. And we sometimes make torn decisions about *important* things; e.g., if you're offered a job you'd love in a city you hate, you might make a torn decision about whether to accept the offer (because a deadline might force you to decide while you're still feeling completely torn).

In sum, torn-D-libertarians agree with determinists about what our lives are *usually* like. They think that we plod through our lives in a roughly Humean/deterministic way, *except when we're making torn decisions*. But as we'll now see, they think that when we make torn decisions, *at least sometimes*, we're L-free.

3.2 The controversial part of torn-D-libertarianism: Consider the following kind of indeterminacy:

---

[3] This, they think, is a *good* thing; for they don't think we should *want* non-decisional actions and no-brainer decisions to be undetermined.

[4] I'm simplifying a bit—because torn decisions and no-brainer decisions are obviously not the only kinds of decisions. E.g., there can decisions in which the agent is leaning toward one option but not sure; and decisions in which the agent has three live options (A, B, and C), and favors A and B over C but is torn between A and B; and so on. But no problems will arise if we just focus on torn decisions and no-brainer decisions.

> A torn decision is *wholly undetermined* at the moment of choice—or, for short, *TDW-undetermined*— just in case the objective moment-of-choice probabilities of the various tied-for-best options being chosen are all roughly even, given the complete state of the world at that moment (and all of the facts about the past) and all the laws of nature, and the choice occurs without any further causal input, i.e., without anything else being significantly causally relevant to which option is chosen. (So if there are two tied-for-best options, then the probabilities are .5 and .5; and if there are three, then they're .333, .333, and .333; and so on.)

This sort of indeterminacy is compatible with various features of the decision being determined. Suppose, e.g., that I'm about to make a torn decision between options A and B. It could be determined that (i) I'm going to make a torn decision (i.e., I'm not going to refrain from choosing); and (ii) I'm going to choose between A and B (and not some third option that I don't like as much); and (iii) the moment-of-choice probabilities of A and B being chosen are both .5 (and not, say, .7, or .3, or 1). The only thing that needs to be undetermined, in order for a decision to be TDW-undetermined, is *which tied-for-best option is chosen*.[5]

The really controversial part of torn-D-libertarianism—the part that anti-libertarians like Sapolsky will want to reject—is captured by the following two claims:

> *The empirical hypothesis (EH)*: At least some of our torn decisions are TDW-undetermined.
>
> *The libertarian thesis (LT)*: If any of our torn decisions are TDW-undetermined, then they also satisfy all of the other conditions for L-freedom—i.e., (i) they're appropriately non-random, and (ii) it's at least partly because they're TDW-undetermined that they're appropriately non-random.

I'll argue in sections 5 and 6 that Sapolsky doesn't say anything in his book to undermine either LT or EH. But in order to set myself up for those arguments, I first need to say a few words about why I think LT is true.

---

[5] There are interesting things to say about cases where the probabilities are neither .5 and .5 nor 1 and 0—e.g., where they're .7 and .3, or .8 and .2. I think that in such cases, there can be *degrees of L-freedom*. But I can't discuss this here.

(To be clear, I don't endorse torn-D-libertarianism because I don't endorse EH. I've argued elsewhere that we have no good reason to *reject* EH; but I also think that we have no good reason to *endorse* it.)

4. A QUICK-AND-DIRTY ARGUMENT FOR LT

In this section, I'll wave my hands at an argument—which I've spelled out in detail elsewhere (2010, 2014)—for LT. The first point to note here is that if indeterminism is true, then there are undetermined events of a certain kind—what we can call *fork-events*—where a *fork-event* is an undetermined physical event that, so to speak, "spits the universe" down one possible path rather than another. The existence of such events follows straightforwardly from the truth of indeterminism.

The second point to note is that if a torn decision is TDW-undetermined, then it *is* a fork event. For example, if I make a torn decision to order chocolate ice cream rather than vanilla, and if this decision is TDW-undetermined, then my conscious decision—the event in which I think to myself *"I'll go with chocolate"—is* the fork-event. I.e., it *is* the undetermined physical event that makes it the case that the universe evolves in a mark-eats-chocolate way instead of a mark-eats-vanilla way. And I think this is just what we want—or *should* want—out of a libertarian kind of freedom. We should want it to be the case that (a) our conscious decisions are the events that settle how the world evolves, and (b) nothing causes us to choose in the ways that we do in these cases. But this is precisely what we get if our torn decisions are TDW-undetermined. E.g., if my decision to order chocolate ice cream was TDW-undetermined, then my conscious decision *was* the event that settled which way the universe would evolve—i.e., whether it would evolve in mark-eats-chocolate way or a mark-eats-vanilla way—and nothing made me choose chocolate over vanilla (in particular, the prior probabilities of the two tied-for-best options being chosen were .5 and .5).

This is an extremely fast version of the argument. A better version of the argument would proceed much more slowly. Moreover, it would include responses to a number of different potential objections. I don't have the space to run through all of this here, but I'd like to say a few words about the following objection:

> *Objection #1*: You say that if your conscious torn decision to order chocolate was TDW-undetermined, then it was the fork-event—i.e., the undetermined physical event that spit the universe down the mark-eats-chocolate path instead of the mark-eats-vanilla path. But that's not right. The undetermined events that spit the universe down this path were *quantum* events. And those events were brutely physical events. So there's no sense in which *you*—the conscious agent—controlled the outcome.

My response to this is that if my torn decision was TDW-undetermined, then the undetermined quantum events in question were *parts* of my conscious decision. Here are two different models of how things might go:

> *The prior-events model*: Some undetermined quantum events occurred, and these events then caused Mark to choose chocolate rather than vanilla.
>
> *The parts model*: Some undetermined quantum events occurred, and those events were the undetermined parts of the undetermined macro-level event—in particular, the *neural* event, i.e., the conscious decision.

If my torn decision was TDW-undetermined, then we know for sure that the prior-events model is *false*; for if the decision was TDW-undetermined, then it was undetermined *at the moment of choice*. So, in this scenario, the relevant undetermined quantum events were *parts* of the neural event—i.e., parts of the conscious decision. Now, at this point, you might respond by saying this:

> *Objection #2*: So what? Even if the undetermined quantum events were parts of the decision, it's still true that those quantum events—and not the conscious decision—were the fork-events.

My response is that objection #2 is confused. Suppose a baseball hits a window and breaks it, and suppose I say this:

> [1] The baseball hitting the window caused the window to break.

Now imagine that someone objects and says that [1] isn't true; rather, they claim, the following is true:

> [2] The relevant subatomic events—i.e., the subatomic events that composed the macro-level event of the baseball hitting the window—caused the window to break.

This objection is confused. [2] is, of course, true. But [1] is also true. Indeed, [1] and [2] say essentially the same thing. And likewise in our case. Let D be my conscious decision. D is a macro-level physical event, in particular, a *neural* event. (Of course, it's not a specific neural firing; it's presumably made up of many neural firings, many neurotransmitter releases, and so on; it's also presumably not a *precisely defined* event, but this doesn't matter.[6]) And like *all* macro-level physical events, D is composed of *quantum* events. Assuming that D was TDW-undetermined, I claim that the following is true:

> [1*] D was the undetermined physical event that spit-the-universe down the Mark-eats-chocolate path instead of the Mark-eats-vanilla path.

The author of objection #2 disagrees; they claim that the following is true instead:

> [2*] The undetermined quantum events that were parts of D are the events that spit-the-universe down the Mark-eats-chocolate path instead of the Mark-eats-vanilla path.

My response: [2*] is, of course, true. But [1*] is also true. And, indeed, [1*] and [2*] say essentially the same thing—in the same what [1] and [2] say the same thing.

So while I haven't fully argued for this here, I think that LT—i.e., the libertarian thesis articulated in section 3.2—is true.

5. SAPOLSKY AND LT

I don't think Sapolsky says anything in his book that gives us any good reason to doubt LT. He says two things that might seem relevant here. First, on pp. 228-230, he asserts—without any argument—that if a decision is undetermined, then it's random. But if my argument for LT is cogent, then this is simply mistaken. Second, Sapolsky argues (pp. 231-237) against the following claim:

> [Harnessing] Conscious agents can "harness" quantum indeterminacies in a way that enables them to have free will.

---

[6] The lack of precision is true of all macro-level events that we denote with ordinary-language event words like 'decision' and 'cab ride' and 'football game'. These events are all made up of quantum events, but there's never a precisely defined set of quantum events that compose a football game or a decision; there are always quantum events which are such that there's no fact of the matter whether they're parts of the game or the decision. And I think it's plausible to suppose that there will always be neural events (neural firings, or neurotransmitter releases, or whatever) for which there's no fact of the matter about whether they're parts of the decision.

I agree that [Harnessing] is false. But torn-D-libertarians don't endorse [Harnessing], and so Sapolsky's arguments against [Harnessing] don't undermine torn-D-libertarianism. It's worth saying a few words about this. First, the language in [Harnessing] is highly suggestive of a *dualist* view. The idea seems to be that a conscious agent is a thing that exists over and above the neural goings on in the given person's head. But torn-D-libertarians reject that view—they're card-carrying *materialists*. Second, the idea behind [Harnessing] seems to be that some undetermined quantum events occur, and then—*after this*—the conscious agent "bubbles those quantum events up into free will". Or something like that. But, again, torn-D-libertarians reject this view. They think that the relevant bunch of quantum events *just is* the conscious decision.

6. SAPOLSKY AND EH

I don't think we have any good reason to endorse EH—i.e., the hypothesis that at least some of our torn decisions are TDW-undetermined. But I've argued elsewhere (2010, 2014) that we also have no good reason to *reject* it. In particular, I've argued that scientists like Libet, Haynes, Wigner, and Tegmark have failed to give us good reasons to reject EH. And I want to argue here that Sapolsky's arguments don't give us good reason to reject EH either.

Sapolsky says only one thing in his book that would undermine EH if it were true. The claim comes in chapter 10, when Sapolsky says that even if there are undetermined quantum events—and, unlike me, he seems convinced that there *are* such events[7]—it's still likely the case that all macro-level events are determined (or close to determined) because it's likely that quantum indeterminacies get "washed out" before we reach the macro-level. But we *know* that this claim is false. If there are undetermined quantum events, then there are at least *some* undetermined macro-level events—namely, macro-level quantum-measurement events. Moreover—and I don't think Sapolsky would disagree with me about this—if there are any undetermined macro-level events, then certain kinds of neural events (e.g., neural firings, neurotransmitter release, and the opening and closing of ion channels) are prime candidates for being events that are (or at least might be) undetermined. This is simply because current neuroscience treats all of these phenomena probabilistically.

[7] I don't think we have any good reason to think that some quantum events are undetermined. The question of whether there are such events is, I think, a completely open question. We just don't have any good reason to endorse either answer to this question.

I don't think Sapolsky says anything else in his book that's even in tension with EH. He argues (correctly) in chapters 3 and 4 that our decisions (like all other neural events) are causally influenced by all sorts of past events—events that occurred just before our decisions or many years before them. But this doesn't undermine EH. For as I've already made clear, if D is a torn decision (e.g., my decision to order chocolate), then the claim that D was TDW-undetermined is fully compatible with the claim that D was causally influenced by the kinds of past events that Sapolsky discusses. Indeed, the claim that D was TDW-undetermined is compatible with the claim that it was *fully causally determined* that at the moment D occurred, I would be making a torn decision, and I would be in the exact mental state that I in fact *was* in—i.e., that I would be torn between chocolate and vanilla. So everything Sapolsky says in chapters 3 and 4 is fully compatible with EH.

And likewise for the chapters on chaos and emergence—because EH doesn't say anything about chaos or emergence. And since EH says only that *some* of our torn decisions are TDW-undetermined, it's fully compatible with the claim that lots of our torn decisions are determined by things that we're unaware of—e.g., by our subconscious beliefs and desires, or by magnetic stimulations to the brain, or by subliminal advertising, or whatever. It's obvious that our decisions can be—and often *are*—influenced (or fully determined) by these sorts of things. But there's no good evidence for the claim that *all* of our torn decisions are determined by these things.

So, again, I don't think Sapolsky has given us any good reason to reject EH. And, to repeat, my own view is that we have no good reason to *endorse or reject* EH.

7. DESIRABILITY

I think that, unlike Sapolsky-freedom, L-freedom is *worth wanting*. I argued for this in my (2010), but I'll just make two points about this here:

1. If you're L-free, then at least sometimes, when you make torn decisions, *you're* the one who settles which tied-for-best option is chosen—i.e., *your conscious decision* is the event that settles this—and nothing *makes* you choose the option that you choose.
2. If you *weren't* L-free, things would be worse for you. For this would presumably be because EH wasn't true, and this would mean that in connection with all of your torn

decisions, which option you choose is caused (or at least causally influenced) by things that you're unaware of and that are external to your conscious reasons and thought.

Given these two points, I think L-freedom is worth wanting.

REFERENCES

Balaguer, M. (2010) *Free Will as an Open Scientific Problem*, Cambridge: MIT Press.

-----(2014) *Free Will*, Cambridge: MIT Press.

Sapolsky, R. (2023) *Determined*, New York: Penguin Press.