University of St. Thomas Journal of Law and Public Policy

Volume 15
Issue 1 *Neuroscience and the Law*

Article 3

October 2021

The Neuroscience of Free Will

Adina L. Roskies

THE NEUROSCIENCE OF FREE WILL

DR. ADINA L. ROSKIES

In this paper I focus on two different ways in which neuroscience has made arguments about the lack of free will or the illusion of free will. One is an argument from determinism, and the other is an argument about the inefficacy of conscious will. I will address both of those different research programs, and I will argue that neither of them actually succeeds in undermining the notion of free will. At the end, I will discuss further the philosophical implications of my views.

I begin by giving an overview of the philosophical landscape of free will. Discussions about free will tend to be framed in terms of the truth or falsity of determinism, and the question of determinism is seen as central to the possibility of free will. There are a number of different philosophical positions that you could have about the relationship of free will and determinism.[1]

Incompatibilism is the position that free will is incompatible with determinism. People who think that we do not have free will, because they think the universe is deterministic and free will is incompatible with determinism are called hard determinists. I include among them Robert Sapolsky, another speaker in the conference of which this paper is a part. However, you might think that free will and determinism are incompatible, but still think we are free because you think the universe is indeterministic. Incompatibilists who think we have free will in virtue of indeterminism. are called libertarians.

However, there are those who can (but need not) accept that the universe is deterministic but believe in free will because they think that free will is compatible with determinism. I think it is worth bearing this in mind, because a lot of the neuroscientific arguments I address ignore this option. I think almost all of them take incompatiblism to be the case and argue for hard determinism. Their general approach is to argue that we can see determinism in the brain, and that if our brains are shown to be deterministic, then that would undermine the possibility of free will. That is really largely because people tend to be naive hard determinists.

---

[1] *See* Adina Roskies, *Neuroscientific Challenges to Free Will and Responsibility,* 10 TRENDS COGNITIVE SCI., No. 9, 2006, at 420.

Now, determinism is a technical term. Essentially, determinism means that if you have a complete specification of the state of the universe at any time, and a complete specification of the laws, that would be sufficient for fixing the state of the universe at all other times.[2] So there is a very specific, and a very stringent, meaning to a claim of determinism.

I want to consider today whether the brain sciences really threaten to prove determinism to be true. There are two, I think, global reasons to think that they cannot. These are called—by me, in Roskies 2008—the problem of fundamental neural indeterminacy, and the other, the problem of exhaustive neural determination. Those are mouthfuls, but to sum it up very quickly: the problem of fundamental neural indeterminacy is that we cannot really tell whether things that look to us to be deterministic (or indeterministic), are really metaphysically deterministic (or indeterministic). The reason is that the only real handle we have on those things is via the operationalization of determinism: predictability.

We tend to think that if we can predict something, it is deterministic, and if we cannot predict it, it is indeterministic. But counterintuitively, we may be able to predict things that are indeterministic, and we may not be able to predict things that are deterministic. To illustrate: chaos, for instance, is unpredictable by definition, even though chaotic behavior can be deterministic. However, if you do not know the governing equations, you cannot predict future states, even if those governing equations are completely deterministic.

On the other hand, there is quantum mechanics which we think of as fundamentally indeterministic. But at the macroscopic level, the physical world can be entirely predictable. When you look at the world, it might look predictable and might look deterministic because it is macroscopic. But that does not mean that it is fundamentally deterministic. Thus, the operational notion of predictability crosscuts the determinism-indeterminism question in a way that does not enable our scientific advances—especially at the neuroscientific level—to tell us much about the true metaphysical status of the world vis a vis the question of determinism.

The second reason I think we will never be able to tell whether the brain is deterministic or indeterministic is that it is a massively interconnected system. You can consider the wiring diagram of the only animal for which we have a complete wiring diagram, and that is the *C.*

---

[2] *See* Nada Gligorov, *Determinism and Advances in Neuroscience,* 14 AMA J. ETHICS 489, 490-91 (2012).

elegans worm.[3] It has 302 neurons, and the interconnections are entirely known, yet incredibly complex, even with only 302 neurons. If you scale that massive connectivity up to a brain with 10^{12} neurons, each of which has a thousand or 10,000 connections, you can understand how incredibly complex our brains must be.

This means that it is impossible to entirely isolate subsystems in the brain. Therefore, if you see something you do not predict, you cannot tell whether that event is the result of true indeterminism, or just the result of deterministic influence from activity somewhere else in the system. In order to actually tell, you would have to have complete knowledge of each neuron in the brain and its interconnections, and I think no matter how good neuroscience gets, it is never going to give us complete knowledge at that level of detail. We will not be able to tell whether what we see is really deterministic or indeterministic. That is why I think neuroscience is not going to settle the question of determinism.

The other main way in which neuroscientists have argued against free will is by claiming to show that conscious will is inefficacious, in other words, that our brain decides things before we do. I will call this the classic Libet argument. The argument basically revolves around a phenomenon that is measured through EEG (electroencephalogram) called the Readiness Potential or RP. This brain potential was first documented half a century ago by Kornhuber and Deecke and employed by Benjamin Libet to argue against the possibility of free will.[4] It has really become a very central issue in the free will debate—people have done a lot of work on the RP trying to understand its role, but the argument has also percolated far beyond neuroscience and has become very significant in philosophical debates about free will, and even in the popular press.

---

[3] *See* Oliver Hobert, Neurogenesis in the Nematode *Caenorhabditis elegans*, *in* WORMBOOK: THE ONLINE REVIEW OF C. ELEGANS BIOLOGY (2005-2018), https://www.ncbi.nlm.nih.gov/books/NBK116086/.

[4] Hans H. Kornhuber & Lüder Deecke, *Brain Potential Changes in Voluntary and Passive Movements in Humans: Readiness Potential and Reafferent Potentials,* 468 EUR. J. PHYSIOL. 1115 (2016); Benjamin Libet et al., *Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initiation of a Freely Voluntary Act,* 106 BRAIN J. NEUROLOGY 623 (1983); Benjamin Libet, *Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action*, 8 BEHAV. & BRAIN SCI. 529, 538–539 (1985).

Let me describe the Libet paradigm. In these studies,[5] the subject is asked to "spontaneously" and periodically move his or her finger or wrist. The subject has an EEG cap on his head, recording brain signals from the scalp. While the subject is spontaneously willing his movements, the movement is being recorded by an EMG machine (electromyogram). At the same time, the subject is also looking at a clock with a rotating dot and is asked to note where that dot is on the clock at the time they spontaneously will their action. Then, retrospectively, they report where that dot was, in order for the scientist to obtain a time for when they consciously willed their action, or what they called W-time.[6]

You record from the central electrode, while the subject repeatedly performs the task. When you take a lot of trials and average those together, time-locked to the motion (EMG signal), you get the RP (Readiness Potential). This is the classic ramp in brain signal that occurs prior to movement.[7] The zero-time is when the EMG machine shows that you moved; the RP begins up to several seconds before the zero-time and has been interpreted as preparatory activity for volitional movement. The Libet experiments, then, aim to measure the time of the consciousness of willing and action relative to the time of the neural signals.

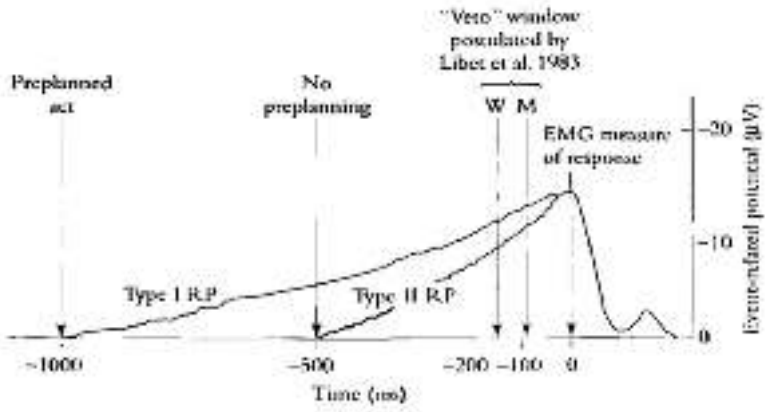
---

[5] Jolyon Troscianko's Website, https://www.jolyon.co.uk/illustrations/consciousness-a-very-short-introduction-2/ (last visited Feb. 23, 2021).

[6] Scott Berry Kaufman, *The Neuroscience of Free Will*, SCI. AM. Mar. 16, 2020, https://blogs.scientificamerican.com/beautiful-minds/the-neuroscience-of-free-will-a-q-a-with-robyn-repko-waller/.

[7] University of St. Thomas School of Law, *Journal of Law and Public Policy Neuroscience and the Law Symposium*, YOUTUBE (Dec. 7, 2020), https://www.youtube.com/watch?v=hwXGJWBAmhQ&feature=youtu.be at 1:47:52 (showing steadily rising readiness potential line with steep drop-off at time 0).

Libet's Experiment[8]



To summarize a lot of work, we are going to consider the Type II RP, which is spontaneously willed actions with no pre-planning. What people have found, and this is a very reliable finding, is that when you average all these traces together, and you look at the time at which people report being aware of willing their movement, it tends to be about 200 milliseconds prior to the movement itself.

That is not problematic in and of itself, but what Libet noticed is that the beginning of those RPs predate the time at which you are consciously aware you are moving, on average, about 200 milliseconds before W-time.[9] This he took to be very problematic because, as he reasoned, if you think of the onset of the RP—where it begins to deviate from baseline—as the decision point, it looks like your brain has already decided to start the movement before you are aware of willing it.[10] He interpreted this as showing that conscious will is illusory or inefficacious.

Libet noted a couple of ways that he thought, perhaps, we could save conscious will. One is that perhaps we could have "free won't."[11] That is, maybe we could veto an action in those 200 milliseconds between becoming

---

[8] Reprinted with permission from William P. Banks & Susan Pockett, *Benjamin Libet's Work on the Neuroscience of Free Will*, *in* THE BLACKWELL COMPANION TO CONSCIOUSNESS 657, 658 (Max Velmans & Susan Schneider eds., 2007). https://doi.org/10.1002/9780470751466.ch52.

[9] William P. Banks & Susan Pockett, *Benjamin Libet's Work on the Neuroscience of Free Will*, *in* THE BLACKWELL COMPANION TO CONSCIOUSNESS 657, 658 (Max Velmans & Susan Schneider eds., 2007).

[10] *Id.*

[11] *See* Benjamin Libet, *supra* note 4.

aware of willing it and actually causing it.[12] But it turns out that does not work at all because vetoing things involves the same kind of preparatory activity, with the same RP/W-time lag. Libet also allows, "in those voluntary actions that are not 'spontaneous' and quickly performed, that is, in those in which conscious deliberation (of whether to act or of what alternative choice of action to take) precedes the act, the possibilities for conscious initiation and control would not be excluded by the present evidence,"[13] suggesting that there might be situations not like the ones that he measured in which the time course is different enough that you might actually be able to exert conscious will. The veto possibility and the longer timescale would still be open under his view, but he thought that, in general—with spontaneous actions at least—actions are not going to be free because they are not able to be consciously willed.

Now, I just want to outline three commitments of this classic view. One is that it is *post-decisional*. That is, the RP—the readiness potential—is neural evidence of a decision to move which has a distinct and measurable onset: the onset of the rise of the RP above the baseline level. Another is that it is *ontologically real*. That is, the RP is a real, causally efficacious signal that initiates action. Finally, it is *not conscious*. That is, the RP precedes subjective awareness of decision. These three commitments lead to the widely accepted doctrine that conscious will is inefficacious, and therefore not free.

Although this view has been very, very popular and has affected people's discussions about free will in many fields, it has not gone without critique.[14] Philosophers and other neuroscientists have critiqued this. For instance, they argue that the experiments are not studying instances of a free decision to move or not, but rather a decision about when and whether to follow instructions given by the experimenter to complete this task.[15] Other critiques say that the simple motor acts that are involved in this task are not the right paradigm for investigations of free will, or that longer time scales are the significant ones, as Libet himself notes, for example, Gollwitzer's

---

[12] *Id.*

[13] *See* Benjamin Libet et al. *supra* note 4.

[14] *See* Karim Fifel, *Readiness Potential and Neuronal Determinism: New Insights on Libet Experiment*, 38 J. NEUROSCI. 784, 785–86 (2018).

[15] *Id.*

implementation intentions occur over longer timescales and seem to have causal efficacy.[16]

One of my favorite criticisms of the classic view is that the task itself fails to measure the time of conscious will. It involves two sub-tasks: one is to spontaneously move your finger, and the second is to report the time of awareness of your intent to move. That itself requires two different components: one is recognizing your own intention to move, and second to index with respect to the clock, an external object, when your intention to move is satisfied, and then report that result. This shift from your inner monitoring to perception of this outer object is an attentional shift, and that shift takes time. In addition, this awareness of your own intention to move is a meta-state, which depends on your already consciously willing something and then must therefore logically follow in time upon it. Thus, W-time does not measure the time of conscious will, but the meta-state that is the awareness of one's own conscious intention.

So, even though people report their W-time to be between when their RP begins to rise and when they actually move, I argue that this is not the time at which they consciously will to move, this is *when they become metacognitively aware of their willing to move*. That is a consciousness of intention, which means that the real W-time, the time of the intention has to precede it, and we do not know exactly where in the time course it would be. This undermines the classic interpretation.

Finally, recent results in neuroscience have also given us other reasons to be really suspect of this way of interpreting the RP. The three criteria that I laid out before are called into question when we look at these other models.

What we do know from years of neuroscientific investigation of the RP is that it is a signal in EEG visible prior to spontaneous voluntary movement, *when you time-lock to movement onset an average over multiple epochs* and usually multiple subjects. When you look at individuals, there is quite a bit of variability in the RP across individuals. It is also very difficult to resolve the RP in single trials because the traces are so noisy that any signal is really masked by the noise. It does seem that the timing of the RP seems to be related to planning, so that if you plan movements, you show an earlier rise in the RP.

---

[16] Peter M. Gollwitzer, *Implementation Intentions: Strong Effects of Simple Plans*, 54 AM. PSYCHOLOGIST 493, 494 –501 (1999).

But there are anomalies. We have noticed that not all subjects exhibit an RP, despite still being able to do the task. And the shape and onset of the RP vary across subjects and across different kinds of trials. More importantly, the RP so far has not proven to be a very good predictor of movement onset; if it were a causal signal leading to movement onset you would expect that you could use it to predict movement.[17]

These are puzzling anomalies. But other work really calls into question the role of the RP as causally effective in causing action. These undermine the idea that the RP reflects a commitment to move. First, it has been shown that it is possible to measure RPs even in cases where people ultimately withhold movement.[18] So you can generate an RP and then not move. RP has also been shown to be present in tasks that do not involve motor components,[19] but involve, for instance, cognitive decisions; RPs are also found to be present in tasks involving decisions not to move.[20] All of this, instead of being consistent with the interpretation that the RP actually is reflective of an initiation of a process of movement, is consistent with the RP being reflective of a decision process itself, but not necessarily a decision process that is the initiator of movement.

Aaron Schurger and colleagues have recently put forth a computational model they call the stochastic accumulator account.[21] This essentially employs a standard diffusion to bound model—which is what psychologists and neuroscientists have been using to model decision processes for years—and shows that an RP-like signal results from averaging

---

[17] *See* Ou Bai et al., *Prediction of Human Voluntary Movement Before It Occurs*, 122 CLINICAL NEUROPHYSIOLOGY 364, 365–70 (2011); *see also* Steven G. Mason & Gary E. Birch, *A Brain-Controlled Switch for Asynchronous Control Applications*, 47 IEEE TRANSACTIONS ON BIOMED. ENGINEERING 1297, 1305–06 (2000).

[18] *See* Matthias Schultze-Kraft et al., *The Point of No Return in Vetoing Self-Initiated Movements*, 113 PROC. NAT'L ACAD. SCI. 1080 (2016).

[19] *See* Prescott Alexander et al., *Readiness Potentials Driven by Non-Motoric Processes*, 39 CONSCIOUSNESS & COGNITION 38, 45 (2016); *see also* R.Q. Cui et al., *High Resolution Spatiotemporal Analysis of the Contingent Negative Variation in Simple or Complex Motor Tasks and a Non-Motor Task*, 111 CLINICAL NEUROPHYSIOLOGY 1847, 1847–49 (2000).

[20] Judy Trevena & Jeff Miller, *Brain Preparation Before a Voluntary Action: Evidence Against Unconscious Movement Initiation*, 19 CONSCIOUSNESS & COGNITION 447, 453–55 (2010).

[21] Aaron Schurger et al., *An Accumulator Model for Spontaneous Neural Activity Prior to Self-Initiated Movement*, 109 PROC. NAT'L ACAD. SCI. E2904, E2905–06 (2012).

many trials, time locked to movement, of a regular stochastic accumulator model of decision. But under this interpretation, the RP does not reflect the time of the decision. The decision time is itself the effect of the bounded accumulator model reaching a certain threshold. That is the decision point.

The model is descriptively adequate. You ask someone to continually, spontaneously move their finger at random intervals and the model essentially replicates the empirical data on waiting time. When you time-lock all of these traces to the point at which someone moves, what you get is exactly what the RP looks like. This is a completely different interpretation of the phenomena that generates the RP.

If you think about the regular classic interpretation, the idea is that there is some kind of baseline level of neural activity and there is a neural decision to initiate movement at the point at which the trace leaves the baseline and then at some point there is enough activity to get the movement to happen. In contrast, the stochastic accumulator model is a *pre-decisional* model. What you see is the evolution of activity prior to decision, and the decision happens at the point at which the threshold is crossed, which is immediately prior to action and not hundreds of milliseconds before the action occurs. So essentially nothing has been decided in what appears to be a build-up to decision, that is just the evolution of decision-related activity, and the decision occurs right before the movement, rather than well before the movement occurs. So, the neural decision to initiate movement immediately precedes the movement itself.

The implication of this stochastic model is that at times where the RP is already detectable, no decision to move has been made yet. The decision thus follows the time at which people report being conscious of willing an action, which is exactly what common sense suggests. It may reflect a buildup of a signal which the agent becomes aware of around that W-time.[22] This is what you might expect if consciousness results from a temporally extended process,[23] as many people have suggested.

---

[22] Elisabeth Parés-Pujolràs et al., *Latent Awareness: Early Conscious Access to Motor Preparation Processes is Linked to the Readiness Potential*, NEUROIMAGE, Nov. 2019, at 1, 7–8.

[23] *But see* Patrick Haggard & Martin Eimer, *On the Relation Between Brain Potentials and the Awareness of Voluntary Movements*, 126 EXPERIMENTAL BRAIN RES. 128, 131–32 (1999); Alexandar Schlegel et al., *Barking Up the Wrong Free: Readiness Potentials Reflect Processes Independent of Conscious Will*, 229 EXPERIMENTAL BRAIN RES. 329, 334–35 (2013).

If one believes the accumulator model, the RP is actually a kind of artifact. It emerges from biased sampling by averaging EEG traces time-locked to action. It does not necessarily reflect an underlying neural process that is present only prior to action because you never end up sampling the cases in which action does not occur. It also implies that there is no physiological significance to that onset point of the RP, the point at which it deviates from baseline. That point is not real in the sense that it is an artifact of averaging and does not reflect a causal process that on any individual trial is related to action. You can get something that looks like an RP signal from time-locking and averaging accumulation processes driven by noise, or by a a variety of other potential, real, small neural signals.

Let's revisit the three commitments of the classical model and see what the stochastic model says about them. First, rather than it being *post-decisional*, it says that it is *pre-decisional*. The RP reflects evidence that contributes to a decision to move. Second, rather than being *ontologically real*, it is *artifactual*. The RP is an artifact of sampling a causally efficacious neural process, which is the decision process itself, but the underlying signal has no measurable onset and may not resemble the RP in individual trials. Third, we really do not know whether it is *conscious* or *unconscious*. The RP may reflect a signal that gives rise to subjective consciousness of a decision or intention, but it may not. Nonetheless, because the W-time precedes the decision time, it does not call into question any of the folk psychological or intuitive positions we may have about free will.

If the onset of the RP is meaningless, relating W-time to it is also meaningless. This completely undercuts the Libet-style arguments for unconscious decisions. According to the stochastic model, the decision in these kinds of Libet-style tasks is actually due to noise. In these cases of spontaneous action, when you have no reason to act at one time versus another, it is perhaps appropriate to let decisions happen using noise or whatever kinds of mechanisms can operate in these conditions. I would argue that the fact that the decision happens due to noise in these kinds of decisions that are set up to have no reasons for choosing one alternative or another, or for choosing to move now as opposed to later, poses little threat to our philosophical conceptions of free will.

The stochastic accumulator model raises questions about how that model interacts with philosophical positions that, I think, are not entirely straightforward. So, depending on your philosophical position, you may or may not think that this accumulator model view of how action is initiated challenges free will. For instance, a Libertarian agent-causation theorist who

thinks agents are special and sort of outside of the causal network of the physical world—that they are uncaused causes—I think they would not be satisfied by this accumulator model because the decision to move *now* is caused by noise: random signals that pass the decision threshold. That probably will not satisfy them as an agent cause. However, it would satisfy many other types of Libertarians, *if* that noise is indeterministic. We do not know whether or not the noise in the brain is indeterministic.

I believe that the stochastic accumulator account would satisfy most compatibilists. A compatibilist could say "Look, in cases where there are not reasons governing our decisions to act, then noise may be exactly the kind of thing that you would want to allow us to make random decisions." But that does not have implications for other sorts of decisions like making decisions for reasons. Moreover, even in the stochastic decision model, the deviation from the baseline that allows one to cross a threshold is a result of intentional and conscious decision. You are putting yourself in the task-set of moving your fingers spontaneously. However, source compatibilists might still be worried, because if the noise is really what is driving the action, then you might argue "Why is this decision in these cases up to me, rather than someone else?" There are different philosophical positions that would cause someone to take this model and interpret its implications for free will differently.

There are other variants of this kind of stochastic model that more clearly reflect volitional processes, such as motivational strength. If one of these were correct, you could have a story in which the threshold crossing events are not entirely due to noise but are causally connected to volition. As long as you have some sort of volitional component, that might be enough to secure agency or free will for decisions, if you are a source compatibilist, or even perhaps if you are an agent-causation theorist.

Many theorists accept that we have agency even if we are not aware of or in control of all the causal influences on our decisions. In fact, social psychology has shown that there are factors that influence our behavior that we are unaware of. Very few people have taken that as an argument that we have no freedom at all, because many accept that it is perfectly compatible with free action that there are still things that bias or causally affect the things we do.

In closing, I want to summarize a couple points. Neuroscience is not going to resolve the determinism question. The classic, or post-decision model, assumes that the RP onset has physiological significance and opens up a space for comparison between onset time and W-time, and thus for

arguments about inefficacious conscious will. However, the family of stochastic models that I introduced can equally well model the RP. Under these models, the classic RP is just a natural consequence of analysis techniques, but it does not entail that in individual trials there is an electrical potential with the characteristics of the RP that signifies action initiation. Rather, the underlying decision processes give rise to decisions that occur approximately when we were aware of them occurring.

I would say that none of these models or neuroscientific results challenges the efficacy of conscious will, even though the model details might affect the way in which we interpret various decisions and the way in which they are related to different philosophical positions. With regard to the ultimate relevance to the law, what I want to say is that I do not think that any of these neuroscientific avenues of exploration are going to succeed in convincing us, or should succeed in convincing us, that we lack free will. I do think that there is plenty of neuroscientific evidence that could be or is relevant to the determinations of responsibility. But I think that they speak to the notion of capacity. I think if we have a capacitarian idea of free will, then we have to have certain kinds of capacities in order to act freely, like capacities to reason, and inhibit our actions, and various other capacities that I think the law clearly recognizes. Then, perhaps we can show neuroscientifically that people's capacities are severely impacted. I think that those are the kinds of neuroscientific pieces of evidence that may actually be much more significant for law. But ways of trying to globally undermine notions of free will and responsibility are all flawed, and I do not really see any avenue for those to impact the law.