compos mentis

Can Neuroscience Comment on Whether We Have Moral Responsibility?

Luke Arend
Bethel University

ACKNOWLEDGMENTS
I would like to acknowledge E.R.R.B., J.A., and C.P. for providing comments and conversation during various stages of the manuscript.

ABSTRACT
This essay discusses the extent to which findings in neuroscience could inform whether or not humans are morally responsible for our actions. First, I argue that the question of moral responsibility maps directly onto the question of free will. Next, I examine two opposing philosophical views on the link between free will and determinism. The incompatibilist position holds that freedom and determinism are mutually exclusive; under this view, we find that science can offer no insights as to whether we have free will, as it can neither prove determinism nor demonstrate freedom. The compatibilist view holds that free will may coexist with determinism; this is accomplished by loosening the metaphysical criterion for freedom. On this view, modern neuroscience can study free will in a limited sense, by conceptualizing free will in terms of the conscious vs. unconscious components of decision-making. I examine several landmark findings of neuroscience, discussing varying interpretations of these results in the context of the greater philosophical tradition. While free will as a metaphysical question is likely to remain untouched by scientific evidence, the findings of neuroscience have certainly proved capable, under the limited compatibilist view, of addressing longstanding popular concepts of conscious will.

KEYWORDS
Moral Responsibility, Free Will, Compatibilism and Incompatibilism, Determinism, Quantum Mechanics, Libet Experiments, Neuroscience

> I am the master of my fate / I am the captain of my soul.
> *William Ernest Henley, "Invictus"*

Free will is dear to us. Arguably one of the greatest motivators of human progress through the ages has been a sense of unlimited self-determined potential—the sheer force of human will rebelling against the maneuvers of fate. Entire civilizations rise and fall based on the principle of unalienable rights owed to every human by mere virtue of their status as a free agent in the world. The entire judicial system hinges on a principle of moral responsibility, and most every religious system in some way acknowledges that we are accountable for our deeds, whether good or ill.

William Ernest Henley captures with chilling resolve the innate human desire for control. Under an alternative interpretation, however, these lines mask an undertone of desperation: backlash to the deep-seated insecurity that we are somehow purely at the mercy of our circumstances. For centuries, philosophers have wondered to what extent, if any, our apparent free agency in the world coexists with the seemingly deterministic structure of everything else around us. Despite the pragmatic need to hold people legally responsible for their actions, and despite the ubiquitous conscious experience that we make hundreds—if not thousands—of freely willed choices every day, a lurking question remains: are we genuinely responsible for our actions, or are we coerced into them by the inscrutable forces of fate? In other words, do we truly have free will? This essay addresses whether findings in neuroscience could answer this question.

First, I will show how the question of moral responsibility directly maps onto the issue of free will. Next, I will discuss two opposing philosophical treatments of free will: compatibilism and incompatibilism.[1] I will argue that under the incompatibilist view, no solid conclusion can be reached as to whether we have free will. Then I will demonstrate how certain findings of neuroscience, when interpreted under the compatibilist view, have indeed nuanced our understanding of conscious free will.

1. The aim of this essay is not to defend either of these two views; my purpose is simply to discuss what each position allows us to conclude about free will.

THE PHILOSOPHICAL PROBLEM

What makes a person 'morally responsible'? Generally, there are two notions linked with the term: "(i) the having of a moral obligation and (ii) the fulfillment of the criteria for deserving blame or praise (punishment or reward) for a morally significant act or omission" (Honderich 2005, 'responsibility'). These notions are linked: praise or blame can be conferred based on whether moral obligation is fulfilled or neglected (Honderich 2005).

Praising or blaming someone for an act either encourages or discourages the repetition of that act in the future. Thus, there is some sense in which moral responsibility presumes that if someone were offered the same choice again—or a sufficiently similar choice—they would have the ability to choose otherwise (Flanagan 1996, 63). This is precisely the link between moral responsibility and free will. If I have free will, then I alone am responsible for selecting any particular action from a set of available actions. Roderick Chisholm explains that free will would mean "each of us, when we act, is a prime mover unmoved. In doing what we do, we cause certain things to happen, and nothing—or no one—causes us to cause those events to happen" (Chisholm 1964, 12).

So, the argument that moral responsibility arises directly from free will is as follows:

> P1: If A causes B to happen, and nothing causes A to do so, then A alone is responsible for B.
>
> P2: If humans have free will, then we cause things to happen and nothing causes us to cause those things to happen.
>
> C: If humans have free will, then we are responsible for everything which we cause to happen.

We can argue the opposite in the absence of free will by the same token. If we do not have free will, then none of our actions are 'uncaused' in the sense above; rather, every decision or action is simply a link in an unbroken chain of deterministic cause and effect. If this is the case, then we cannot be held morally responsible for our actions any more than a car can be held morally responsible for a car accident. A person who commits a murder, for instance, does not actually make this decision but is coerced into it—they are simply the murder weapon in

the hands of unseen precedent causes over which they have no control.

So, we have established the following connection: an agent is morally responsible for its actions if and only if it is a free-willed agent. By showing that these two concepts go hand in hand, the original issue—whether we have moral responsibility—is reduced to an equivalent question: do we have free will? This is the question we will seek to answer going forward.

As previously alluded to, the issue of free will is closely related to that of determinism. Aptly put, determinism holds that "all events without exception are effects—events necessitated by earlier events" (Honderich 2005). If this is the case, the whole universe is, per William James' famous imagery, a fixed 'iron block' of causality in which the future is equally immutable as the past.

Traditionally, philosophical discourse on the relationship between free will and determinism has fallen into two camps. Incompatibilism holds that if determinism is true, humans cannot have free will; on the other hand, compatibilists hold that we can accept both free will and determinism, most often by arguing that our actions can be 'caused' without being 'coerced' (Honderich 2005).

THE INCOMPATIBILIST APPROACH

It appears that the incompatibilist could easily have their answer to the free will question by showing that determinism is true: if all is determined, then we have no ability to choose otherwise. Ergo, free will does not exist. The chemical and electrical processes of the brain are no exception to the rigid laws of cause and effect; all thoughts, words, and deeds alike are meticulously orchestrated by the same physical dynamics which govern the motion of planets and the toppling over of a sequence of dominoes.

The problem with this is that no scientific experiment could ever show that determinism is true: any such effort is doomed to stop short of certainty due to the problem of induction. No matter how regularly we observe determinism to hold in any particular instance, logic does not warrant the conclusion that it is an inviolable universal law. By its nature, scientific induction is only capable of falsifying the thesis of determinism—never verifying it.[2] Thus, the incompatibilist

2. One might object to this, saying that empirical evidence in favor of determinism can be amassed to the point at which determinism is so highly probable that one may reasonably believe that things are so. Naturalism—the predominant scientific worldview—indeed takes this to already be the case. I have no issue with this. A high degree of evidence-based belief, amounting to *practical certainty*, is distinct from absolute metaphysical certainty about determinism. My

view cannot truly say with certainty that we have no free will; it can only hold that we do not have free will if determinism is unequivocally true, the latter being a fundamentally unprovable presupposition.

On the other hand, if the incompatibilist were to find that determinism is not true, this still would not prove positively that we have free will. This is because a system could be indeterministic in two different ways: (i) due to genuine agency or (ii) due to pure chance.

Grant for a moment, despite the enormous practical difficulties, that we can set up a neuroscience experiment to show that the brain is indeterministic. Imagine we can somehow isolate a brain (and any necessary surrounding environment) in such a way that it is totally undisturbed by outside activity. Further, entertain for a moment that—contrary to the predictions of quantum mechanics—we can measure the precise state of the entire 'brain-system', down to the very last particle, without disturbing it in the slightest. Absolute determinism dictates that any system, when set up just so, will evolve in time in a completely predictable way; whatever conditions there are at the onset provide a fixed description of what happens at all other times.[3] All it would take to falsify determinism would be to set up two separate trials starting with identical systems, and, after a fixed period of time, discover that something different resulted in each case.

---

argument concerns only the latter.

3. For instance, given the complete 'state' of a flying projectile (a description of both its position and momentum), we can predict where it is going *equally well* as where it came from. This 'both-ways' predictability is a hallmark of any deterministic system, resulting from causal symmetry. From a purely physical standpoint, the cause-effect relationship traveling forward in time is indistinguishable from the effect-cause relationship traveling backward in time. The perceived *direction* of causality is dictated by nothing more than the direction in which the arrow of time is classically defined (that in which entropy increases as a system evolves).

   Under the traditional (Copenhagen) interpretation of quantum mechanics, such predictive symmetry does not hold: prior to observation, a system is described by a 'wavefunction' or probability distribution; after observation, as a single particle. We can only make probabilistic predictions about how the wavefunction will 'collapse' upon observation. As state information is discarded in the collapse, such predictions can only be made forward in time across this event. Without a one-to-one mapping of possible states from each moment to the next, the system no longer undergoes an invertible transformation through time. This interpretation of quantum mechanics paints a fundamentally indeterministic picture of the world in which causality, as we know it, is violated.

Now let us perform our fantastical experiment. We set up identical brain systems as aforementioned, and, as hoped, we observe different outcomes in each case! We collect our Nobel prize. We have produced incontrovertible evidence that decisions—including moral ones—are not deterministic.[4] However, our demonstration of indeterminism is still a far cry from empirical proof for the positive existence of free will. In this fantastical experiment, two possible explanations remain for why the brain-system had the ability to choose differently. First, perhaps we witnessed true agency—the brain exercised its free will and chose differently in each trial. However, it could equally be the case that no free will was involved: the difference arose due to pure chance. The traditional interpretation of quantum mechanics suggests that the universe is fundamentally probabilistic[5]—a particle's behavior, upon observation, seems to be dictated by nothing but chance. In our experiment, then, perhaps the brain-system (neurons, atoms, particles, and all) simply evolved differently in each trial due to randomness; the differential "decisions" each occurred by dumb luck. Of course, this would be greatly removed from anything resembling genuine free will—this type of agent would bear no more moral responsibility than one which could only make moral decisions by rolling dice (Honderich 2005).

Thus, the incompatibilist reaches an impasse: determinism cannot be proved true, and even falsification of determinism leaves the free will question unresolved. If we accept the incompatibilist view, there is no definitive way to show whether we have free will, and, consequentially, no finding—in neuroscience or otherwise—could decide whether or not people are morally responsible for their actions.

4. A thoughtful reader might point out that an experiment involving far less than an entire brain-system could serve to falsify determinism—witnessing even a single poorly-behaved electron would do the trick. I stand by the brain example because it gives a fairer chance to an experiment which not only disproves determinism but positively demonstrates free will in a human-like agent. Falsifying determinism is theoretically possible. Yet, even this highly idealized experiment fails to show that free will exists, as I shall presently argue.

5. Or, at least, the Copenhagen interpretation holds that *our predictions* about a physical system can only be fundamentally probabilistic—even with perfect knowledge of the initial conditions. An interesting essay might explore whether this simply equates to a sort of *epistemological* indeterminism, distinct from any metaphysical commitment.

THE COMPATILIBILIST APPROACH

There is an alternative view, however. Compatibilism—as the name implies—seeks to harmonize free will and determinism: Owen Flanagan calls it "the position that the reality of voluntary action is fully compatible with an analysis of such action as caused" (Flanagan 1996, 57). For compatibilists, determinism need not spell out the death of free will; in fact, free will could be argued for or against whether or not determinism is true.

However, if the answer to the free will question is not based on establishing determinism or non-determinism, where can we look? Neuroscientists have looked to gain insight by turning directly to the supposed 'seat of agency': the brain. Questions of determinism are all but ignored in the neuroscience literature, which instead often focuses on analyzing the causal relationships between unconscious neural activity, conscious decision making, and resulting actions. An empirically workable definition of free will only requires that conscious decisions cause actions, not that those decisions themselves are metaphysically uncaused (Carruthers 2007, 198). Thus, when neuroscientists ask whether we have 'free will', they are perhaps asking whether we have conscious will: are decisions ultimately made by neural processing which occurs at the conscious or subconscious level? This approach—a form of compatibilism—allows neuroscientists to seek out empirical evidence for or against 'free will' while sidestepping the gaping metaphysical problem of determinism.

Much debate over free will in the neuroscience community has arisen in the wake of a set of landmark experiments by Benjamin Libet (1985). In short, Libet found that conscious awareness of a spontaneously willed decision was preceded by an unconscious neurological readiness potential (RP) predicting the 'willed' motor action. From this result, he argues that conscious free will does not operate the way we often envision it, namely, making high-level selections of action from a wide range of options. Rather, our subconscious brain generates actions while conscious will merely has the final 'veto-power' to permit or prevent the consummation of those actions (Libet 1985, 551).

Libet strikes a nuanced balance by suggesting that we are not consciously responsible for our thoughts—only the resulting actions. On the one hand, he preserves naturalistic determinism by acknowledging the causal structure of an underlying RP which initializes intentions and precedes thought. At the same time,

moral responsibility can be conferred due the fact that the behavioral output is modified by a conscious decision.[6]

Alfred R. Mele is skeptical about this interpretation, challenging the association between the RP and intention. In another experiment by Libet, subjects were instructed to prepare to flex their fingers at a given clock time, and then to consciously "veto the developing intention/preparation to act" instead of following through with it (Libet 1985, 538). Here a ramp-like RP was still found, but instead of fully developing into the moment of action, it dropped off "about 150-250 ms before the preset time," suggesting that the conscious veto prevented the RP from being carried through into motor action (Libet 1985, 538). Mele argues from this that the RP cannot represent an intention to act: here, the RP is present while the subject has an intention not to act all along, and it is illogical that a subject could intend both to act and not to act at the same time (Mele 2006, 193). Thus, Mele finds Libet to be mistaken in identifying the RP as the intended action which is vetoed; furthermore, he notes that such interpretations can "quickly get out of hand" when applied nonchalantly to the nuanced philosophical issue of free will (Mele 2006, 197).

Mele instead proffers that the generation of an act can be broken down into a multi-part process that begins with an unconscious urge (the RP), yet is "directly initiated" by intention on the conscious level (Mele 2006, 199). Thus, the RP does not represent a decision or intention (in the sense that the act is set in motion at the subconscious level and can only be vetoed by conscious will); it instead represents an 'urge' which then may or may not be initiated by the will (Mele 2006, 199). Mele's alternative explanation seems to show, at the very least, that Libet's finding is far from a definitive ruling either for or against free will.

6. One might object to the claim that moral responsibility can be conferred in this case, saying that my earlier argument only equated moral responsibility with free will in the strict metaphysical or causal sense—not with mere conscious will. I reply that even if our definition of 'free will' only entails conscious will, affirming free will for humans implies a coherent notion of moral responsibility. Recall that moral responsibility is the conferral of praise or blame in order to encourage or discourage similar future behavior. At least from the psychological viewpoint of the agent, this is an effective and sensible strategy as long as the decision is made at the conscious level. Thus the agent can be held morally responsible in a meaningful way. If the decision is made at the subconscious level, however, then a notion of moral responsibility collapses: it would seem rather torturous to punish someone for a subconscious decision. Thus, the equivalency between moral responsibility and free will holds for the compatibilist and incompatibilist alike, albeit in slightly different senses.

Interestingly, Patrick Haggard remarks that the common notion of free will—while an important aspect of our folk psychology—is incompatible with modern neuroscience due to its implication of mind-body dualism (Haggard 2005, 291). He affirms Libet's interpretation that 'free choice' is driven by unconscious processing, pointing to an experiment in which Ammon and Gandevia (1990) used transcranial magnetic stimulation—without the subject's awareness—to bias a subject's choice to flex one wrist or the other. Despite significant findings such as these, Haggard notes that conscious will has still not received nearly as much research attention as phenomena relating to conscious perception (Haggard 2005, 291). Clearly, much remains to be explored.

CONCLUDING REMARKS

Can neuroscience comment on whether we have free will and moral responsibility? It depends on a metaphysical choice of perspective. Under the incompatibilist view relating free will and determinism, it is impossible for science to ever establish definitively whether we are free. On the other hand, many neuroscientists take the compatibilist approach, studying conscious free will as a scientific question separate from the metaphysically intractable issue of determinism. On this more limited view of agency, the findings of neuroscience have certainly proved capable of commenting on free will and, consequently, moral responsibility. The Libet experiments, while not closing the case either way, are a promising early step in this investigation; at the very least, the fact that their interpretation has been so hotly contested is a testament to their significance. Though no findings have yet resolved whether we have free will, these results carry implications which have unsettled longstanding folk concepts of agency and conscious will. We can reasonably expect that the findings of neuroscience will continue to do so.

REFERENCES

Ammon, K. and S. C. Gandevia. 1990. "Transcranial magnetic stimulation can influence the selection of motor programmes." *Journal of Neurology, Neurosurgery, and Psychiatry* 53 (8): 705-707.

Carruthers, Peter. 2007. "The illusion of conscious will." *Synthese* 159 (2): 197-213.

Chisholm, Roderick M. 1964. "Human Freedom and the Self." *The Lindley Lecture*, University of Kansas, April 23, 1964.

Flanagan, Owen. 1996. "Neuroscience, Agency, and the Meaning of Life." In *Self Expressions: Mind, Morals, and the Meaning of Life*. New York: Oxford University Press.

Haggard, Patrick. 2005. "Conscious intention and motor cognition." *Trends in Cognitive Sciences* 9 (6): 290-295.

Honderich, Ted, ed. 2005. *The Oxford Companion to Philosophy*. New York: Oxford University Press.

Libet, Benjamin. 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action." *The Behavioral and Brain Sciences* 8 (4): 529-566.

Libet, Benjamin. 1999. "Do we have free will?" *Journal of Consciousness Studies* 6 (8-9): 47-57. Reprinted in Kane, Robert, ed. 2004. *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Mele, Alfred R. 1997. "Strength of Motivation and Being in Control: Learning from Libet." *American Philosophical Quarterly* 34 (3): 319-332.

Mele, Alfred R. 2006. "Free will: Theories, analysis, and data." In Pockett, Susan, William P. Banks, and Shaun Gallagher, eds., *Does Consciousness Cause Behavior?* Cambridge, Massachusetts: MIT Press.

Wegner, Daniel M. 2003. "The mind's best trick: how we experience conscious will." *Trends in Cognitive Science* 7 (2): 65-69.

Wegner, Daniel M. 2004. "Précis of The illusion of conscious will." *Behavioral and Brain Sciences* 27 (5): 649-692.