Free will is not a testable hypothesis

Abstract

Much recent work in neuroscience aims to shed light on whether we have free will. Can it? Can any science? To answer, we need to disentangle different notions of free will, and clarify what we mean by 'empirical' and 'testable'. That done, my main conclusion is, duly interpreted: that free will is not a testable hypothesis. In particular, it is neither verifiable nor falsifiable by empirical evidence. The arguments for this are not a priori but rather are based on a posteriori consideration of the relevant neuroscientific investigations, as well as on standard philosophy of science work on the notion of testability.

1. Introduction

Do we have free will? Label by *Testable* the view that this can be tested empirically, where by 'empirically' I mean third-person investigations such as those of neuroscience and cognitive psychology. Testable is endorsed by many neuroscientists[1] and, as we will see, also by many philosophers, at least implicitly. Nevertheless, I will argue, free will is not an empirically testable hypothesis and thus Testable is false.

I will depart from much of the existing literature in two ways. First, I focus exclusively on the metaphysical existence question. Thus, I say nothing about moral responsibility or about free will's relation to normative theorizing more generally. Second, my arguments are not a priori. Rather, they are based on a posteriori consideration of relevant empirical investigations. I do not have in mind hypothetical discoveries that might somehow tell us whether the universe as a whole is deterministic, but instead the actual neuroscientific research programs underway now into human decision-making. What is at stake is whether disagreement about the existence of free will can ever be resolved empirically. The originality of the argument here will lie in making the case for untestability without, so to speak, fixing the philosophical deck in advance, for instance by a priori dismissing incompatibilist free will as unintelligible, or as unacceptable because it contradicts methodological naturalism.

To start with, I will understand free will along classic incompatibilist lines, namely as, roughly speaking, human agents having some capacity to influence the world independent of prior physical causes. Free agents are taken to be in control of what they do (including deciding) in a way that distinguishes their free behavior from random behavior. I will consider various refinements as we go along, concentrating on those that are most favorable for Testable and showing that even then Testable fails.[2] Then I will turn to the (much easier) case of compatibilism.

The word 'testable' is ambiguous: do we mean testable now or testable also in the future? I will argue that free will is not testable now (and so has not yet been verified or falsified) and will not be testable in the near future. With respect to the distant future I will conclude, with respect to human investigators and given the entrenched background assumptions on either side, that whether free will will ever be testable is an open question but that there is no particular reason to expect that it will be.

It will help to consider the two faces of empirical testing separately: first verification and then falsification. After that, I will examine the very notion of testability more closely.

2. Is free will verifiable?

What feasible *empirical* findings would establish to a believer in Testable that free will exists? My answer is: ‘none’, or at least none that are feasible now or in the near future, and arguably even in the distant future.

To see why not, begin by clarifying exactly what a confirming piece of evidence would be. Initially, one might think of something along the lines of a physical action that an agent reported having willed and yet that had no detectable prior physical causes. Now, presumably there would be some proximate physical causes even in this case. For instance, the raising of a hand would be preceded by nerve impulses travelling to the muscles in the arm – if not, then the raising of the hand would be a miracle quite independent of free will. So when speaking of ‘prior physical causes’ here, we have in mind neurological events in the brain. But this still needs to be refined further – for if the agent was conscious at all, there would continuously be many neurological events in their brain. A conscious, freely willed decision itself, though, would presumably lead to, or perhaps be constituted by, a specific new set of neurological events in the brain. This set of events would in turn physically cause the hand to be raised. So the key evidence, finally, would be that this new set of neurological events has no detectable prior physical causes.

Further refinement is required, for by ‘caused’ do we mean fully determined or merely probabilistically caused? Here, the latter would be sufficient for verification. That is, if no fully determining unconscious causes of the conscious decision-making could be found, that would be evidence of free will. If prior neurological events made an action more likely but did not determine it – in other words if they were merely probabilistic causes of it – that would still leave some part of the causal pie, so to speak, to free will.[3] The prior neurological events would be akin to mere ‘influences’ or ‘dispositions’.

In summary so far, the relevant situation would be that, at the moment of conscious decision, a set of neurological events begins the chain of physical causation that leads to the hand being raised. The verification condition would then be: that we cannot detect any prior physical causes that fully determine this set of neurological events.

However, would such a situation indeed establish the existence of free will? No. After all, in effect we are in this epistemic position already, and indeed always have been, for there are of course many actions for which neuroscience does not yet have a complete causal explanation. But merely finding something unexplained in itself proves nothing – we find unexplained things all the time. What would be required instead is finding something unexplain*able*, which is a much taller order. For how could we ever establish this stronger claim, namely not only that we currently have no complete explanation but also that we will not in the future?

The only scientific[4] route to a demonstration of free will is therefore via a proof of *absence*. In particular, what must be proved is the absence of fully determining prior physical causes of the activity in the brain that corresponds to conscious decision-making.[5] It is this epistemological demand that is the fundamental difficulty because, given our incomplete (to put it mildly) knowledge of the brain, it is very hard to rule out the relevant causes being merely undiscovered rather than absent.

How have absences been established in other areas of science? Examples do abound. For example, radioactive emissions have been established not to be caused by anything external, and hidden variables ruled out (according to most) as a cause of quantum entanglement. More generally, it is routine for trials and experiments to establish that one thing does not cause another, for instance that the MMR vaccine does not cause autism.

The exact methods used to establish these absences vary case by case. But they do have one crucial factor in common, namely that, so to speak, they are cases of *total* absence – hidden variables simply do not exist; there are no external causes at all of radioactive emissions; and the control groups in the MMR studies are known to have received no MMR vaccine at all. By contrast, in the free will case no one thinks that neurological events in the brain do not occur at all. Rather, the requirement is to establish that they do not fully determine a particular subset of effects, namely conscious decision-making. This is a much more difficult task. It is relatively easy to establish complete causal isolation for a system in order then to rule out the influence of external prior causes – think of a radioactive atom in a sealed box, or a plant in a sealed vessel. But there is no such total causal isolation in the case of a living brain making decisions, nor is it feasible to create it. Similarly, there is no prospect of establishing a control group with no neurological events.

As a result, in the free will case the prospects of establishing the necessary absence are dim. It seems that only a complete causal knowledge of the brain's operation, such that every physical event at the neural level is fully explained, would suffice. None of the usual short cuts for establishing absences is available. This makes the task infeasible, at least in the near future and probably for rather longer too. Whether the required level of knowledge will ever be feasible for human investigators, and thus whether free will will ever be empirically verifiable, seems to me an open question.

Nothing in the above arguments is a priori. There is no appeal, for instance, to the claim that incompatibilist free will must rest on non-physical causation, which in turn is necessarily undetectable by any empirical investigation committed to methodological naturalism (footnote 5). Nor is the possibility of autonomous conscious decision-making held to be incoherent with a physicalist worldview, or the very notion of incompatibilist free will held to be unintelligible. Rather, the difficulties are a posteriori and practical. Even if we do not adopt philosophical presuppositions that, as it were, rule out the possibility of free will in advance, still even then verification seems infeasible.

3. What would falsify free will?

Now turn to the flipside issue. What feasible empirical findings would establish to a believer in Testable that free will does *not* exist? My answer is again: 'none'. Accordingly, free will is not empirically falsifiable. I will proceed in two stages: in this section, I consider what evidence would be needed for falsification. Then, in the next section, I argue that such evidence is unobtainable. Along the way, we will see why arguing against Testable is not arguing against a straw man.

Start by noting that many have suggested that falsification is already to hand, or at least soon may be. The famous series of experiments, starting in the early 1960s, by the neurologist Benjamin Libet (1985) claim to demonstrate that volitional action is initiated unconsciously. In particular, an electric readiness potential is reliably detectable in the brain more than 500 milliseconds before a physical action, but subjects report the moment of consciously deciding to perform the action to be only 200 milliseconds before. Similar results have now been replicated many times. Is the readiness potential a physical cause of the subsequent physical action? To assess this, we need to know whether it increases the action's probability, and to assess that in turn, we need to know how often a readiness potential was *not* followed by the action.[6] The original Libet experiments only tracked cases where the action did occur, so cannot answer this latter question. However, later studies, using a variety of methods, have done so. And sure enough, it has been found that unconscious probabilistic causes do exist. It is accordingly possible to predict with greater than chance accuracy whether an action will occur before the subject consciously decides whether to perform it. In one study, such physical causes were detected a full four seconds before participants were aware of making a choice (Soon et al 2013); in another, up to 10 seconds (Soon et al 2008).[7]

Do these results falsify free will already? No – because the causation they establish is only probabilistic. Label by *neuroprediction* the prediction of an action on the basis of the physical causes detected in a subject's brain before their conscious decision. The accuracy of these neuropredictions has been much less than perfect. In both the Soon studies, for instance, they were correct only roughly 60% of the time – barely better than the 50% accuracy achievable by pure chance. The highest accuracy reported in any study is 80% (Fried et al 2011). There is plenty of room for a subsequent conscious decision to 'change' matters after the neuroprediction is made. Indeed, many participants in the Libet studies reported doing exactly this. Accordingly, there is still room to claim a role for free will.

Several writers have argued convincingly for a similarly negative verdict (e.g. Mele 2009, 2013, Nahmias 2014, O'Connor 2009, Mylopoulos and Lau 2014, Maoz et al 2014). They offer many additional reasons. One of these is that so far all the neurological studies have concerned simple, trivial actions such as raising a finger or flexing a wrist. But much, perhaps most, of our interest in free will concerns its role in drawn-out and serious matters such as deciding whether to accept a new job or buy a new house. (Kane 2008, for example, holds that 'self-forming actions', in his view the most fundamental exercises of free will, occur only in a subset of such drawn-out cases.) No current neuroprediction experiment speaks to them.

There are several other objections besides, both conceptual and methodological. First, many philosophers dispute whether the operation of free will should always be associated just with conscious choice. Second, to the eye of a philosopher of action, Libet-type experiments are very crude. For instance, no distinction is made between decisions and desires or urges, even though much philosophical theory deems such a distinction highly relevant. Third, it might be that conscious decision-making is more likely to play an independent causal role in choices between morally non-equivalent actions, such as in cases of resisting temptation. It is questionable whether Libet-type results should be extrapolated to these other cases without further evidence. Meanwhile, methodological objections include whether participants' introspective timing reports are reliable, and whether the readiness potential is indeed the neural signature of action initiation (footnote 7). All such objections, if accepted, would render free will's falsification by neuroscience experiments more difficult and so would *help* this paper's thesis. But even just the first reason from earlier, namely that the causation is only probabilistic, will prove sufficient for our purposes. Again, the overall strategy is to show that Testable fails even under the most friendly assumptions.

So far, then, the evidence is not enough to falsify free will.[8] But what really matters for our purposes, and what so far appears to have been largely neglected in the literature, is the more fundamental question: what evidence *would* do so?

The natural answer is: *perfect* neuroprediction. That is, if we could predict physical actions with 100% accuracy before subjects reported consciously deciding on them, then, assuming the subjects' reports are accurate, free will would be falsified. All incompatibilist positions accept that free will cannot exist if human action is fully determined by prior physical causes. The best possible falsifying evidence would therefore be perfect prediction, which in practice means perfect *neuro*prediction.[9] If it were achieved, it would close the probabilistic escape hatch above, for if an action were already 100% determined then there would no longer be any room for free will to intervene – or at least, there would be evidence that no such intervention ever actually takes place.

Determinism would falsify incompatibilist free will, but we do not know whether determinism is true. The point here is that perfect neuroprediction would be the necessary evidence to demonstrate it, or rather, to demonstrate that free will is having no effect. Perfect neuroprediction is the relevant criterion for us because it is the way to demonstrate the impotence of free will *empirically* by means of neuroscience, rather than by appeal to extra-empirical metaphysics.

Some would dispute that even perfect neuroprediction is sufficient. For instance, imagine a Kantian utopia in which all agents acted strictly according to duty. Then their actions would be perfectly predictable and yet nevertheless freely chosen. The underlying point is that it matters on what basis a successful prediction is made. If it is on the basis of the right kind of reason, then that need not falsify free will. Here is a different caveat: Mele (2013) argues that if there were only a short time lag between the unconscious signal and the conscious awareness of intent, this could still be consistent with the agent having freely willed their decision yet only becoming consciously aware of having done so part way through the process. Genuine falsification must conclusively demonstrate that conscious deliberation is impotent in decision making, and perfect neuroprediction is not enough for this. What would be required in addition would be either more detailed knowledge of the exact causal sequences in the brain (see also Mele 2009, 156-8), or else a much longer time lag between signal and conscious awareness.

Again, if these objections were accepted, they would *help* the paper's thesis by making falsification more difficult. However, once again I will assume the most favorable case for Testable, namely that perfect neuroprediction is indeed sufficient for falsification. Testable will not survive even then.

These objections to sufficiency do suggest a different danger though. In particular, are existing incompatibilist theories of free will committed to Testable at all? If not, then arguments against Testable are arguments against a straw man, at least in the case of philosophers. There are relatively few explicit discussions of this issue in the incompatibilist literature, whose focus has typically been on other matters.[10] Nevertheless, upon examination I think it is clear that at least some of the leading incompatibilist theories are indeed committed to Testable, or so I will argue now. In particular, they are committed to accepting that perfect neuroprediction (or at least some cases of it) would falsify free will. Moreover, to my knowledge few philosophers have explicitly *denied* that empirical evidence could bear on whether free will exists, i.e. few have explicitly denied Testable. Thus, a rejection of Testable would indeed break fresh dialectical ground.

Now consider various incompatibilist positions in more detail. Start with Nozick's (1981) event-causal account. According to it, roughly speaking, freely willed actions are those caused by reasons of the right sort. If perfect neuroprediction were possible based purely on factors that theory deemed reasons of the wrong sort, or not reasons at all, that would suggest that the right kind of reasons are causally inefficacious, which in turn by Nozick's own lights would imply a falsification. Libet-style readiness potentials, for instance, would seem clearly to be an example of the wrong kind of factor; thus, perfect neuroprediction on the basis of them would indeed falsify free will. Similar remarks apply to other event-causal accounts, such as (Kane 1996). They apply to the range of non-causal ownership accounts as well (e.g. Ginet 1990, Pink 2004). These too must assume that a person's choice is not wholly determined by prior factors, or at least not by some kinds of prior factors, which in practice means that they could not accommodate perfect neuroprediction, and certainly not if the neuroprediction were on the basis of something like Libet's readiness potentials. And Ginet, for one, implicitly endorses Testable when he criticizes O'Connor's theory precisely because, Ginet charges, it is *not* empirically testable (1997, 96).

What of agent-causation views (e.g. Chisholm 1976, O'Connor 2000, Clarke 2003)? Would perfect neuroprediction falsify these theories too? If not, then I would classify them as being committed to the existence of free will a priori, and thus not to be committed to Testable in the first place. What discussion there is of this in the literature has, not surprisingly, focused on the case of an indeterministic world. In particular, imagine that the results of agents' free choices just so happen exactly to replicate the frequencies of those actions that would have been generated anyway by purely physical laws. This seems conceivable, and so in an indeterministic world there is no contradiction between the exercise of free will and the observation that all events are consistent with scientific law. O'Connor concedes that this renders free will empirically untestable (1995, 195): "it seems that it is impossible in principle, for us ever to know whether any events are produced in the manner that the agency theory postulates, because such an event would be indistinguishable from one which was essentially random."[11] So does Clarke (1993, 99): "there is no observational evidence that could tell us whether our world is an indeterministic world with agent causation or an indeterministic world without it." [12] But Pereboom disagrees (2003, chapter 3). He argues that such a coincidence is statistically implausible and therefore that the lack of any observed law-violations is already strong empirical evidence against free will. In effect, he claims that agent-causalist free will is empirically testable, and moreover has been tested and has come out badly. The logic behind this claim is intricate though. It turns on whether we have any well-founded beliefs regarding what probability distribution we should expect to describe the outcomes of freely willed actions. I will not discuss that here. The relevant point for us is that agent-causalists are unpersuaded, and not clearly irrationally so. Dialectically, the empirical evidence that Pereboom cites has therefore *not* been decisive because it is unable to transcend the two sides' differing background assumptions. As we will see, such a situation is the very definition of untestability.

So, it is arguable that agent-causal theories should indeed be read as rejecting Testable. On the other hand, the general doubt that perfect neuroprediction would cast on indeterminism suggests that it might nevertheless constitute evidence against even agent-causal free will. But either way, other incompatibilist theories of free will clearly can be reasonably construed as committed to Testable, so overall Testable is not a straw man.

Finally, turn to an entirely different issue, nothing to do with straw man worries: never mind whether perfect neuroprediction is sufficient for falsification; is it even necessary? The motivation to ask comes from the famous objection, going back at least to Hume, that mere physical indeterminism does not in itself ensure room for free will. Indeed, by making physical events beyond the full control of any agent, indeterminism arguably even tells against it. So why should lack of perfect neuroprediction, and the implication that we cannot rule out neural events being indeterministic, be deemed an escape route for free will? The answer is that the issue at hand is *epistemic*. Suppose that in fact there is no free will. How could we demonstrate that? Answer: by perfect neuroprediction. But lacking that, *we* cannot rule out that any imperfection of prediction is due to the operation of free will rather than to mere physical indeterminism. Therefore, epistemically speaking, in order to rule out free will, perfect neuroprediction is indeed necessary.[13]

4. Why free will is not falsifiable

I conclude that perfect neuroprediction is in practice the only route by which free will might be empirically falsified. It is clearly necessary for that and, in this section, I will assume that it is also sufficient, thereby again for the sake of argument considering only the case most favorable for Testable. The crucial question then is: will perfect neuroprediction ever be feasible? I answer ‘probably not’, and thus that free will’s falsification will probably never be feasible either. There are two powerful reasons why not.

First, successful models of complex systems are usually probabilistic, and this feature is persistent. In particular, this (so far) well describes neuroscience. In Roskies’ words (2006, 420): “The picture that neuroscience has yielded so far is one of mechanisms infused with indeterministic or stochastic processes. Whether or not a neuron will fire, what pattern of action potentials it generates, or how many synaptic vesicles are released, have all been characterized as stochastic phenomena in our current best models.” The timing of electrical ‘spikes’ emitted by cortical neurons, for instance, has long been apparently stochastic. Usual practice has therefore been to create higher-level constructs such as ‘average firing rate’ in order to ensure successful modeling. Moreover, probabilistic principles are increasingly being incorporated into models of many decision-making (and other) cognitive processes (Chater and Oaksford 2008). Data in psychology too “almost invariably show probabilistic rather than deterministic causation” (Baumeister 2008, 67). Models of monkey decision-making, in some ways more advanced than those of humans because greater instrument penetration of monkey brains is possible, also remain universally probabilistic (Roskies 2014). Overall, there is no induction in the science towards more deterministic models; quite the contrary. Accordingly, there is no induction towards perfect neuroprediction.

Moreover, some recent findings suggest that “the brain seems to incorporate deliberately an element of randomness into its decision-making processes. The neural mechanisms that generate choices during resource acquisition, for example, seem to reflect an added ‘bonus’ for probabilistic exploration of new environments and new alternatives … Our world, including human cognition, is shot through and through with probabilism” (Newsome 2014). The explanation is that decision-making is in this way made more efficient: “the occasional random choices perform the same creative function as occasional random mutations in the genome – they allow exploration of a much larger space of possibilities than would be encountered by simple deterministic processes” (Newsome 2014 – see there for further references). That is, there is positive reason to think that at the neurological level the processes themselves – not just our models of them – are indeterministic. In which case, perfect neuroprediction will never be achievable.

Second, many complex systems are *chaotic*. In the manner of the butterfly effect, in such systems deterministic predictions require indefinitely fine-grained knowledge of initial conditions. In practice, this means that perfect prediction is infeasible; probabilistic modeling is forced upon us permanently. At the moment, no one knows whether the relevant brain systems are actually chaotic in this way. But to assume that they are not would amount to no more than an act of faith, unsupported by any trend in the science.[14]

To be sure, these considerations do not yet *prove* that we will never be able to falsify free will empirically. However, they do show two things: first, that it is not falsifiable now, nor will it likely be in the near future. And second, that there is no particular reason for optimism that it will become falsifiable as science advances. Rather, it seems likely that it will not.

Two further points: first, if free will is defined to obtain only if the decisions that cause actions are conscious, then free will could be falsified by showing that consciousness plays no role in action generation. Again, perfect neuroprediction would be necessary for that, this time prediction of decisions by exclusively non-conscious (as opposed to pre-conscious) neurological processes, assuming that they could be identified. But for similar reasons to those just discussed, it seems dubious that such perfect prediction could ever be achieved by humans.

Second, would demonstrating the impossibility of perfect neuroprediction be tantamount to demonstrating the absence of fully determining causes? If so, the very reason for free will's unfalsifiability would imply also its verification. But the arguments against the possibility of perfect neuroprediction applied only to our neurological *modeling* of the brain, not to the brain itself. In order to falsify free will, we need to show that physical causes in the brain completely determine action, and the only way we can do that is if our neurological models predict perfectly. On the verification side, we need to show that there are no fully determining physical causes. But the fact that our best models are stochastic is not in itself enough to establish this metaphysical claim. This is because that would require the further inference that the causes left unmodeled cannot be physical – and this inference needs independent justification. In other words, in order to show deterministic causation, perfect prediction by models is sufficient. But in order to show *lack* of deterministic causation, *im*perfect prediction by models is *not* sufficient – because now we need to say something also about unmodeled causes too. None of this rules out the possibility of verification a priori: as per section 2, there exist methods in science for establishing the required absence of full determination by physical causes. But it does underscore one reason why verification of free will is so difficult.

5. What is testability?

So far, we have been making do with a somewhat intuitive understanding of testability, but we should now examine this notion more closely. Among other things, doing so will clarify the way in which any claim of untestability is inevitably relativized to background assumptions. It will also thereby clarify how this paper's particular claim of untestability bears on tests that, even though falling short of outright verification or falsification, are still evidentially relevant to some degree.

Duhem and Quine famously taught that no claim is testable on its own but rather only in conjunction with various auxiliary (i.e. background) assumptions (Duhem 1954, Quine 1951, 1955). Furthermore, according to Quine, testability is always a matter of degree: any hypothesis, no matter how deeply entrenched within a belief system, might conceivably be jettisoned in the face of some observation or other. Others have disputed this (Laudan 1990);

I return to the issue below. In any case, in practice the epistemic statuses of a hypothesis and of associated auxiliary assumptions are often highly asymmetric. As a result, a verification (or falsification) will redound to their credit (or blame) correspondingly asymmetrically. It is this asymmetry that saves a claim about testability from triviality: ‘everything is testable’ yes, but ‘everything is equally easily testable’ no. In terms of Quine’s own holistic metaphor, a hypothesis and auxiliary assumption may differ in how close each is to the periphery of our web of belief and thus may differ in their susceptibility to empirical revision.

What matters for our purposes is whether an empirical test can adjudicate a dispute between competing hypotheses. The question of testability becomes the question of whether a dispute is *empirically decidable*.[15] If both sides agree on relevant auxiliary assumptions, an empirical test may indeed be decisive. If not, it may not be. As Duhem himself emphasized, there can be no single ‘crucial experiment’ in isolation; rather, an experiment can only ever be crucial given suitable agreement about auxiliary assumptions.

It is a truism of under-determination that many different sets of beliefs are consistent with the same set of observations. Logically speaking, it is thus contingent which particular sets of beliefs participants actually do hold. The take-home message is that testability is inevitably *relativized* to the particular sets of beliefs that competing sides happen to have. There is no absolute fact of the matter about whether any particular dispute is testable; rather, it will be testable given some participant background beliefs but untestable given others.

With this philosophy of science groundwork in place, return to the case of free will. What is the relevant difference between the two sides’ background assumptions? It does not concern what would constitute either verification or falsification: the analysis in the previous sections is not at issue. Rather, the difference shows in the response to the fact of empirical undecidability, i.e. to the fact of neither verification nor falsification being achieved. In effect, each side puts the burden of proof on the other. Thus, if there is no proof either way, believers take that to leave belief in free will undamaged, while skeptics take it to leave *dis*belief in it undamaged.

Generally, to resolve disputes in science it is necessary that relevant auxiliary assumptions are themselves independently testable (Sober 1999). But in the case of the free will debate, this is just what is not possible – hence the empirical intractability. This is because the relevant auxiliary assumptions are themselves rooted in deeply held philosophical commitments, such as sympathy to naturalism or its opposite, or the epistemic primacy to be given – or not – to third-person empirical evidence. And such disagreements are of course not simply decidable empirically. So, we are stuck.

In summary, given the background beliefs of the competing research programs, current evidence cannot cause either side to revise their foreground belief about free will.[16] Accordingly, the free will dispute is not empirically decidable.

I want to emphasize that this relativization to the two sides' background beliefs does not render free will's untestability therefore somehow mitigated compared to other cases of untestability – because *all* untestability claims are so relativized.

Turn at last to the issue of degree. Imagine, for instance, that scientists achieved 99% accurate neuroprediction. Although therefore short of perfection, might this be enough to shake even advocates' faith in free will, and moreover, crucially, might it be significantly more feasible to achieve than 100% predictive perfection? Similarly, might even some imperfect level of brain mapping, and thus of proof of absence, be enough to cause skeptics to change their minds, and again, crucially, be more feasible to achieve? In which case, the argument runs, the mere infeasibility of *absolute* verification or falsification is not yet enough to establish the untestability thesis.

In reply: free will's in-principle testability is already implied by the in-principle revisability of all beliefs. So what matters is whether evidence *sufficiently* strong to sway beliefs will become *actually* available. And the record suggests not. For example, even 80% or 90% predictive accuracy would clearly have minimal dialectical effect. It is not clear exactly what level would make a big difference, but at a minimum likely something close to the aforementioned 99%.[17] Yet, for the reasons discussed in section 4, achieving even that or anything close seems fanciful. Rather, the general picture is that, by the time a piece of evidence has become feasible to obtain, it has lost its ability to change core beliefs; and by the time a piece of evidence would be strong enough to change core beliefs, it has become infeasible to obtain. Given different auxiliary assumptions it might have been different, but given the ones the two sides actually have we are stuck. Evidence sufficiently strong to sway beliefs is not available, and moreover there seems little prospect of it becoming so.

To sum up: it is not that the existence of free will is 'in principle' forever empirically untestable – but then arguably nothing is ever untestable in this absolute sense. Rather, free will is untestable in the sense that matters: given the auxiliary assumptions that the two sides are actually committed to, decisive evidence is unobtainable.

6. Compatibilism

So far, we have only considered free will as understood by incompatibilists. What about compatibilists? All agree that whether we have a case of compatibilist free will is frequently empirically testable – it is often easy to observe, for instance, whether an agent is making a decision under coercion.[18] Many popular defenses of free will against alleged or prospective neuroscientific falsification have in effect defined free will in compatibilist terms (e.g. Nahmias 2015). That is indeed an effective defense. However, for our purposes it is also unsatisfactory in that it begs the question against incompatibilism.

Overall, assessing the testability of free will is a twofold procedure: first, is compatibilism or incompatibilism correct? Second, if compatibilism is correct then free will is testable; but if not then it is not. But what of the first stage – is the compatibilism-incompatibilism dispute empirically decidable? Surely not, given the range and depth of substantive philosophical commitments that would determine the implications of any empirical findings. If so, then the overall untestability thesis remains intact. Neither the truth of incompatibilist free will, nor the prior compatibilism-incompatibilism dispute, is decidable empirically.

7. Conclusion

Free will is not empirically verifiable because we cannot feasibly rule out that otherwise unexplained actions are due merely to so far undiscovered (non-agential) physical causes. And it is not empirically falsifiable because the perfect neuroprediction needed for that is infeasible.

There are many other issues surrounding free will that, by contrast, certainly are susceptible to empirical enlightenment and that are being actively investigated: What is the role of conscious intention in human decision-making? What influences our attributions of free will, and of moral responsibility? What role did attributions of, and subjective experience of, free will play in our evolutionary past? What role do they play now in our social and ethical discourse and theorizing? Perhaps skeptics of free will may also find in the science raw material for filling out deflationary or error theories. Nevertheless, regarding the central metaphysical question – does free will exist? – empirical testing cannot help us. Or at least it cannot help us now or in the near future, and perhaps ever.

Acknowledgements
I would like to thank an audience at Birkbeck philosophy department, Anna Alexandrova, and anonymous referees at this and another journal for useful feedback.

References

Baumeister, R. (2008). ‘Free Will, Consciousness, and Cultural Animals’, in J. Baer, J. Kaufman, and R. Baumeister, (eds.), *Are We Free? Psychology and Free Will*. New York: Oxford University Press.

Chater, N., and M. Oaksford (eds) (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press.

Chisholm, R. (1976). *Person and Object*. LaSalle: Open Court.

Clarke, R. (1993). ‘Toward a credible agent-causal account of free will’, *Nous* 27, 191-203.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*, trans. from 2nd ed. by P. W. Wiener, Princeton, NJ: Princeton University Press. Originally published as *La Théorie Physique: Son Objet et sa Structure* (Paris: Marcel Riviera & Cie., 1914).

Fischer, J. (1995). *The Metaphysics of Free Will*. Oxford: Blackwell.

Fried, I., R. Mukamel, and G. Kreiman (2011). ‘Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition’, *Neuron* 69, 548–562.

Gazzaniga, M. (2011). *Who’s in Charge? Free Will and the Science of the Brain.* New York: HarperCollins.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Ginet, C. (1997). ‘Freedom, responsibility, and agency’, *Journal of Ethics* 1, 85-98.

Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. (2008). ‘Three Freedoms, Free Will, and Self-Formation. . .’, in N. Trakakis and D. Cohen (eds.), *Essays on Free Will and Moral Responsibility*. Newcastle: Cambridge Scholars Publishing, pp. 142-162.

Knobe, J. (2014). ‘Free Will and the Scientific Vision’, in E. Machery & E. O.’Neill (eds.), *Current Controversies in Experimental Philosophy*. Routledge.

Laudan, L. (1990). ‘Demystifying Underdetermination’, in *Scientific Theories*, C. W. Savage (ed.), (Minnesota Studies in the Philosophy of Science, vol. 14), Minneapolis: University of Minnesota Press, pp. 267–297.

Libet, B. (1985). ‘Unconscious cerebral initiative and the role of conscious will in voluntary action’, *Behavioral and Brain Sciences* 8, 529–566.

Libet, B. (2002). ‘The Timing of Mental Events: Libet's Experimental Findings and Their Implications’, *Consciousness and Cognition* 11: 291–99.

Maoz, U., L. Mudrik, R. Rivlin, I. Ross, A. Mamelak and G. Yaffe (2014). ‘On Reporting the Onset of the Intention to Move’, chapter 10 in Mele (2014), pp184-196.

Mele, A. (2009). *Effective Intentions: The Power of Conscious Will*. New York: Oxford University Press.

Mele, A. (2013). ‘Unconscious decisions and free will’, *Philosophical Psychology* 26, 777–789.

Mele, A. (ed) (2014). *Surrounding Free Will: Philosophy, Psychology, Neuroscience*. Oxford: Oxford University Press.

Murray, D., and E. Nahmias (2014). 'Explaining Away Incompatibilist Intuitions', *Philosophy and Phenomenological Research* 88.2, 434-467.

Mylopoulos, M. and H. Lau (2014). 'Naturalizing Free Will', chapter 7 in Mele (2014), pp123-141.

Nahmias, E. (2014). 'Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences', in W. Sinnott-Armstrong (ed.), *Moral Psychology, vol. 4: Freedom and Responsibility*. Cambridge, MA: MIT Press.

Nahmias, E. (2015). 'Why We Have Free Will', *Scientific American* January pp77-79.

Newsome, W. (2014). 'Neuroscience, Explanation, and the Problem of Free Will', in Walter Sinnott-Armstrong (ed) *Moral Psychology: Volume 4: Free Will and Moral Responsibility*. Cambridge, MA: MIT Press.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

O'Connor, T. (1995). 'Agent Causation', in T. O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, New York: Oxford University Press, 173–200.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press

O'Connor, T. (2009). 'Conscious Willing and the Emerging Sciences of Brain and Behavior', in N. Murphy, G. Ellis, and T. O'Connor (eds.) *Downward Causation and the Neurobiology of Free Will* (Springer Verlag), pp173-186.

Pereboom, D. (2003). *Living Without Free Will*. Cambridge: Cambridge University Press.

Pink, T. (2004). *Free Will: A Very Short Introduction*. Oxford: Oxford University Press.

Quine, W. V. (1951). 'Two Dogmas of Empiricism', in *From a Logical Point of View*, 2nd Ed., Cambridge, MA: Harvard University Press, pp20–46.

Quine, W.V. (1955). 'Posits and Reality', in *The Ways of Paradox and Other Essays*, 2nd Ed., Cambridge, MA: Harvard University Press, pp246–254.

Roskies, A. (2006). 'Neuroscientific challenges to free will and responsibility', *Trends in Cognitive Science* 10, 419-423.

Roskies, A. (2014). 'Monkey Decision Making as a Model System for Human Decision Making', chapter 12 in Mele (2014), pp231-251.

Schurger, A., M. Mylopoulos, and D. Rosenthal (2016). 'Neural antecedents of spontaneous voluntary movement: a new perspective', *Trends in Cognitive Sciences* 20.2 77-79.

Sober, E. (1999). 'Testability', *Proceedings and Addresses of the American Philosophical Association* 73: 47-76.

Soon, C. S., M. Brass, H. J. Heinze, and J.-D. Haynes (2008). 'Unconscious Determinants of Free Decisions in the Human Brain', *Nature Neuroscience* 11, 543–545.

Soon, C. S., A. H. He, S. Bode, and J.-D. Haynes (2013). 'Predicting free choices for abstract intentions', *Proceedings of the National Academy of Sciences* 110.15, 6217-6222.

[1] Libet (2002, 292) gives a list.

[2] Two cases of this already: first, several writers argue that free will should be seen as something that comes in degrees (e.g. O'Connor 2009). But I will focus on the simpler issue of whether there exists free will to any degree at all, arguing that even that is untestable. Second, it is disputed by some that we must require free behavior and random behavior to be distinguishable. If we do not then free will is much harder to test (section 3 below), but generally I will assume that we do.

[3] This does assume that the unaccounted-for portion cannot legitimately be put down to unknown causes other than free will – an important condition. I return to this point at the end of section 4 below.

[4] As with 'empirical', by 'scientific' and cognate terms I mean third-person investigations.

[5] In practice, in the eyes of many even a proven absence of this sort would not be deemed enough for verification. Most scientists would undoubtedly further demand a detailed positive theory of free will, which made novel empirical predictions, could be related to other scientific theories, and so on. Perhaps imposing this further demand is equivalent to imposing an a priori commitment to methodological naturalism. But my concern in this paper is that empirical verification of free will may still be impossible even without such a commitment.

A proven absence of prior physical causes also does not rule out some non-physical cause (other than free will). But, although this is undoubtedly a logical possibility, I do not think that either side of the free will debate takes it seriously, in which case it is not dialectically significant. If we did take it seriously that would only make verification still harder, so ultimately supporting this paper's thesis.

[6] Even then, to establish causation we would need to rule out the readiness potential and subsequent physical action being merely two independent effects of a common cause. But for our purposes we may ignore that possibility, since it would still imply the action having been caused by *some* prior unconscious event.

[7] That said, this interpretation of the neurological evidence is not unanimous. Recently, Schurger et al (2016) have argued that the build-up in the readiness potential is more akin to background noise than to any signal or 'decision' to act. If so, falsification is even further away than will be argued in the text.

[8] Various other types of empirical evidence have been claimed to threaten free will too, such as some results from social psychology or the symptoms of some clinical mental disorders. But these too are unconvincing falsifiers – see, e.g., O'Connor (2009, 179-180) for discussion.

[9] Perhaps a breakthrough in the metaphysical interpretation of quantum theory would generate a consensus that the universe as a whole is deterministic, thereby rendering incompatibilist free will empirically falsified without any need for neuroscience at all. But: first, this currently seems an unlikely deus ex machina. Second, it would not adjudicate between compatibilism and

incompatibilism, thus still leave open which of those is correct, and thus still leave open whether we have free will. And third, at least one incompatibilist, namely O'Connor (2000 chapter 6), denies that it would falsify free will anyway, on the grounds that a deterministic quantum mechanics at the micro-level still would not rule out emergent – and possibly indeterministic – causal powers at higher levels.

[10] Mele is one exception (2009, 157-8). Roughly speaking, he argues that falsification would follow from any demonstration that our psychological decision-making system is deterministic. I think, in turn, that the best *evidence* for this system being deterministic would be perfect prediction at the appropriate level. Thus, I take this to endorse our criterion of perfect neuroprediction.

[11] On the other hand, elsewhere O'Connor seems in effect to urge the opposite attitude, i.e. to *endorse* Testable: "It is an *open empirical* question … whether and when the basic capacity to choose, to which philosophers give most of their attention, is present and regularly exercised in a way necessary for true freedom of choice" (2009, 185, my italics).

[12] This line of argument could also be used against the verification condition for free will suggested in section 2. The thought is that even if an absence of fully determining physical causes of volitional actions were established, nevertheless those actions could still be ascribed to chance. If so, that would render free will unverifiable as well as unfalsifiable.

[13] Suppose no one ever has any understanding of why they are doing what they are doing – as suggested, for instance, by Gazzaniga (2011). As an anonymous referee points out, this claim would, if accepted, falsify free will without perfect neuroprediction being necessary, and regardless of whether determinism is true. But for it to establish Testable, Gazzaniga's claim would need (as per section 5) to be accepted by all sides, and I do not think that prospect is likely – for instance, non-naturalistic defenders of free will certainly haven't accepted his claim so far.

[14] Neither this nor the previous reason implies that the universe itself is indeterministic. The indeterminism is either purely with respect to our best *models* or else only at the neurological level.

[15] Indeed, arguably empirical testing is *always* a comparative matter, being between a proposition and an alternative. Because this alternative may not always be the proposition's simple negation, strictly speaking testability is therefore a property of a problem, i.e. of a proposition plus contrast, rather than of a proposition singly (Sober 1999). In our case though, the competing propositions are indeed 'we have free will' and its negation, so we may elide this technicality without loss, as in the paper's title. The points developed in the text, about testability's relativization to auxiliary assumptions, are unaffected.

[16] In Bayesian terms, each side's degree of belief in whether free will exists will not be altered significantly.

[17] The degree of belief in free will is not related linearly to the accuracy of prediction. Thus, 60% accuracy clearly does not imply 0.6 degree of belief in (lack of) free will, and neither need 99% accuracy imply 0.99 degree of belief.

[18] Two nuances: First, if compatibilists are viewed as being committed to the causal efficacy of conscious decision-making, then perhaps empirical investigation could threaten even compatibilist free will – although that would merely confirm that it is testable. Second, I am ignoring science fiction scenarios, such as aliens secretly manipulating our every action so that what we ordinarily think are uncoerced decisions are in fact anything but. Even then, the untestability of compatibilist free will would further require that this alien conspiracy could never be discovered by us. But in any case, ultimately, as the text explains, the paper's overall thesis stands or falls regardless of whether compatibilist free will is testable or not.