Free will and the Unconscious Precursors of Choice

Forthcoming in *Philosophical Psychology*

This is the author's copy, which may differ from the final print version

Abstract. Benjamin Libet's empirical challenge to free will has received a great deal of attention and criticism. A standard line of response has emerged that many take to be decisive against Libet's challenge. In the first part of this paper, I will argue that this standard response fails to put the challenge to rest. It fails, in particular, to address a recent follow-up experiment that raises a similar worry about free will (Soon et al. 2008). In the second part, however, I will argue that we can altogether avoid Libet-style challenges if we adopt a traditional compatibilist account of free will. In the final section, I will briefly explain why there is good and independent reason to think about free will in this way.

Introduction

Over the past few decades, a number of scientists and commentators have argued that traditional notions of free will and moral responsibility have to be abandoned or, at least, substantially revised in the light of empirical evidence. Most famously, perhaps, Benjamin Libet has argued that experiments concerning the timing of conscious intention show that the brain "decides" to initiate simple motor actions before we become aware of the corresponding intention to make the movement. Virtually every part and aspect of Libet's argument has been scrutinized, and his interpretation of the data has been criticized by both scientists and philosophers. From this, a standard line of response has emerged that many take to be decisive against Libet's challenge. In the first part of this paper, I will argue that this standard response fails to put the underlying worries about free will to rest. I will argue, in particular, that this response is not decisive against a charitable interpretation of Libet's challenge, and we will see that it fails to address a more recent follow-up experiment that has strengthened and radicalized the challenge (Soon et al. 2008). This empirical challenge depends, however, on a particular conception of free will, according to which acting with free will requires the power to initiate. In the second part, I will propose an alternative account of free will that is in line with a traditional compatibilist approach. On this view, free will requires the absence from constraints (such as coercion and interference) and the ability to do otherwise, but not the

power to initiate. I will argue that this conception of free will is immune to any kind of Libet-style challenge. In the final section, I will then briefly explain why there is independent reason to think about free will in this way. Taken together, this will give us ample reason to endorse the proposed account of free will. But first of all, I shall provide a summary account of Libet's seminal experiment.

Libet's challenge

Libet has conducted various experiments concerning the timing of movement and conscious awareness. The challenge to free will is based, primarily, on an experiment with the following design (Libet et. al. 1983, Libet 1985, 1999, and 2004, for instance). Subjects are asked to sit down and to perform a simple and predefined movement (flexion of finger or wrist) when they feel like doing so: they are instructed to let the urge (wish or intention) to perform the movement "come on its own" (however, within a certain period of time). During this, EEG measurements are taken from the midline of the subject's scalp (roughly, over motor and premotor areas) and EMG signals are taken from the muscles that are involved in the movement. At the same time, subjects are looking at a clock-like device and they are instructed to remember and report the position of the clock's revolving dot at the time when the conscious choice to move is made (Libet referred to this conscious event variously as conscious *urge*, *wish*, *choice*, or *intention*). This allowed the experimenters to determine the timing of the onset of the conscious event, which Libet called time W (for *will*). It was already known that voluntary movements are preceded by a brain potential that arises from motor and premotor areas and that is called the *readiness potential* or *RP*, for short (Kornhuber & Deecke 1965). In Libet's experiment, the onset of the RP, obtained by EEG over the scalp, occurred on average 550ms before the onset of the movement. The main result of the experiment was that the onset of the RP also preceded W, the reported time of conscious will, on average by 350ms.

Libet concluded from this that the "initiation of a spontaneous voluntary act begins unconsciously" (1985, p. 529). We can distinguish here between the following two claims. Firstly, the data shows, according to Libet, that "the brain 'decides' to initiate or, at least, to prepare to initiate the act before there is any reportable subjective awareness that such

a decision has taken place" (p. 536). Secondly, Libet concluded that this shows that the conscious intention does not initiate the movement, because it could *initiate* the movement only if it preceded or, at least, coincided "with the onset of the specific cerebral processes that mediate the act" (p. 529).

However, on the basis of further experiments and some observations, Libet also argued that there is evidence for what he called *veto control*. During debriefing, some subjects reported that they had sometimes chosen not to act on the urge to move, thereby aborting the action. And in similar experiments in which subjects were asked to first prepare and then abort the movement, EEG signals were obtained that resemble the recording of RPs. This suggests, according to Libet, that the time window of about 200ms between W and the onset of the movement allows subjects to consciously veto and abort the performance of the action.

Libet's main conclusion was that our conception of free will needs to be revised in the light of this. In particular, he thought that the experimental results are incompatible with the philosophical and commonsense conception, according to which free will requires the power to initiate actions consciously. But the results are compatible with a more modest conception, according to which free will consists in the ability to consciously control by way of vetoing—we don't have free will, but we have *free won't*.

Response strategies

Libet's suggestion that we should rest content with the view that our powers of free agency are limited to veto powers has not found many followers, for the simple reason that this amounts to the admission that we do not have *genuine* free will. Given this, it is no surprise that Libet's claims have met a great deal of resistance, scrutiny, and critique. We can distinguish, very broadly, between two types of criticism. The first concerns the experimental methodology, and the second concerns Libet's interpretation of the data and his conclusions concerning free will and conscious choice.

The most common objection to the experimental methodology has been that the measurement of W (the timing of conscious will) is flawed, or at least inaccurate, and that this is problematic as even small inaccuracies in the measurement of the timing can undermine the conclusions due to the relatively short time frame of the whole process.

But this objection was not successful and the basic results of the experiment are now widely accepted as empirical facts (for a review see Klein 2002, for instance). Moreover, the mentioned methodological problems concerning the measurement of W have been eliminated in a recent experiment. John-Dylan Haynes and his colleagues (Soon et al. 2008) have conducted a Libet-style experiment that measures also the relative times of conscious choices and of the associated neural processes. But they have used different methods. In particular, the clock-like device has been replaced by a screen on which subjects view a stream of letters that updates every 500ms. This measurement method is obviously more coarse-grained, as the unit of measurement is almost as long as the whole RP process in Libet's experiment (550ms). Despite this, Haynes and colleagues have been able to reproduce the basic result, as they have traced back neural activity that is predictive of the choice for up to 10s before the time of conscious choice.

My focus here, however, will be on the second type of response that concerns Libet's interpretation of the data and the conclusions about free will. We can distinguish here further between two responses, or two response strategies, that have been particularly prominent. I shall call the first the narrow strategy, and the second the broad strategy. Together, they form what has become a standard response to Libet's challenge. Before we turn to those two response strategies, let me add some terminological notes. Firstly, I shall use the term 'P-intentions' in order to refer to proximal intentions that immediately precede the performance of basic actions (such as bodily movements) and whose content is specific to the performed type of act (such as the intention to perform "that movement now"). And I shall use the term 'D-intentions' in order to refer to distal intentions that are further removed from the execution of the action and that are, usually, less specific (such as the intention to "go to the movies on the weekend").[1] Secondly, I shall assume that decisions (or choices) just are formations of intentions. Accordingly, I shall use the terms 'P-decisions' and 'D-decisions' to refer to the formation of P-intentions and D-intentions, respectively. Thirdly, I shall use phrases such as 'unconscious neural processes' in order to refer to neural processes that are not correlated

[1] The content of both P- and D-intentions can be more or less specific, and sometimes D-intentions are as specific as P-intentions. The crucial difference consists in the temporal distance from the execution and in the fact P-intentions are assumed to facilitate online control, whereas D-intentions must be stored in order to guide future behavior.

or associated with any of the subject's conscious states and that are below the level of consciousness because they are inaccessible to conscious awareness (as opposed to being not accessed by consciousness).

The narrow strategy

This response consists, basically, in an alternative interpretation of the data. According to Libet's interpretation, subjects become aware, at W, of a decision that has *already been made*—or of a decision-making process that is *already underway*, as it were. On this view, the RP stems from one single process that is accompanied by a subjective awareness only after a significant delay. According to an alternative interpretation, subjects become aware of an urge or desire *in response to which* a conscious P-intention is formed. On this view, the RP stems from a process that consists, basically, of two stages. In the first stage, an unconscious urge or desire is generated. In the second stage, the subject consciously decides to act in response to becoming aware of that urge or desire. This shows, according to the narrow strategy, that the empirical evidence does not support the claim that the brain *decides* to move before the conscious choice is made, because the evidence is consistent with the hypothesis that subjects consciously decide to act in response to becoming aware of an urge or desire.

Libet addressed this response and he dismissed the alternative interpretation as "ad hoc speculation" (1985 p. 535). Since then, this response has been defended and developed in detail by the philosopher Alfred Mele (1997 and 2009, for instance). Mele emphasizes the need for conceptual clarity and rigor. Conceptually, it seems clear that there is a robust difference between motivational states, such as desires and urges, and executive states, such as choices or formations of intentions. This favors the alternative interpretation of the data over Libet's, which conflates the two types of mental states. More importantly, perhaps, further research on the RP has provided convincing empirical evidence for the view that the early and late components of the RP stem from distinct neural sources (for a review see Shibasaki & Hallett 2006, for instance). This shows that the alternative two-stage interpretation is not altogether an ad hoc speculation, as Libet claimed.

The broad strategy

This response concerns the significance of the experimental result and Libet's argument against free will. According to Libet, the experiment shows that we lack free will with respect to the timing of simple movements. But he argued also that this undermines free will in general. The main idea here is that all decisions, no matter how spontaneous and no matter how deliberate, eventually have to "work their way into" the neural mechanisms that initiate the relevant motor actions (1985, p. 530). Simple motor decisions are the final gates that all decisions have to pass through, as it were, in order to result in overt behavior, and so it seems that we cannot have free will unless we have free will with respect to final motor decisions.

The broad strategy targets this argument. It is usually put forward by way of the following observations concerning the experimental setup (compare Keller & Heckhausen 1990, Flanagan 1992, pp. 136-38, and Zhu 2003, for instance). In the Libet experiment, subjects are asked to perform a certain type of movement within a certain time frame, and they agree to follow these instructions. In particular, they consciously decide to follow the instructions, and they are, usually, able to execute this commitment. Given this, it seems clear that their actions during the experiment are influenced and guided by D-intentions that are formed consciously at the beginning of the experiment. It seems clear, in other words, that the conscious decisions made at the beginning of the experiment do, somehow, "work their way into" the motor control system. These D-decisions do not determine the exact timing of the movement. But they guide its execution in two important respects: they result in the requested type of movement and they are executed within the requested time frame—they are, as far as that goes, clearly efficacious. Given this, it seems clear that Libet's challenge to free will does not generalize to all decisions—at least not without further evidence or argument.

Moreover, it seems that the Libet challenge concerns only choices that are relatively insignificant. Most of our decisions have a what-component and a when-component—we decide what to do and when to do it. Usually, what-decisions are more significant, because most when-decisions are made in order to determine how to implement a certain what-decision. When-decisions are in this sense subordinate to what-decisions. This is highlighted by the fact that the when-component is initially often left

unspecified or vague. One may, for instance, decide to see a doctor, leaving the exact timing open and subject to external constraints. Moreover, moral responsibility is usually attached to what-decisions, rather than when-decisions. If one robs a bank, for instance, one will be held responsible for doing *that*, and not for the particular timing of the deed. In any case, it seems that one is praise- or blameworthy for a when-decision only if one is praise- or blameworthy for the implemented what-decision. Given, then, that Libet's challenge concerns only when-decisions, it concerns only decisions that are usually in the service of more important decisions what to do.

A decisive response? The narrow strategy

Both the narrow and the broad strategy raise important issues, and they clearly show that Libet's conclusions are questionable. But they are, as I will now argue, not decisive. Let us first return to the narrow strategy. At the heart of this response is the claim that there is no good reason to think that the brain *decides* when to move before the corresponding conscious choice is made. A first point to note here is that Libet himself put "decide" in scare quotes in the often quoted passage (1985, p. 536, quoted above). This suggests that not even Libet thought that the brain is, literally, making an unconscious decision as opposed to, say, generating an unconscious urge or desire. Given this, it would be uncharitable to assume that Libet's challenge is based on the claim that subjects only become aware of a decision that has already been made unconsciously. But how should we interpret the challenge?

One possible interpretation says that the challenge is based on the claim that the early-RP corresponds to a neural process that *causally determines* the occurrence of the conscious P-intention and the subsequent movement. This claim, however, is in tension with Libet's own veto experiments. If the early-RPs in veto cases stem from the same neural activity as the early-RPs in the paradigm experiment, then it cannot be the case that this activity causally determines the conscious P-intentions and the movements. Moreover, it is quite clear, I think, that this is not what Libet had in mind.

Another interpretation says that the challenge is based on the claim that the early-RP corresponds to a neural process that *will* result in the conscious P-intention and the movement, *unless* the agent decides to veto. This interpretation supports the following

two claims. Firstly, it suggests that the conscious P-intention is not the causal initiator of the movement, because it occurs too late—because it neither precedes nor coincides "with the onset of the specific cerebral processes that mediate the act" (ibid., p. 529). This suggests, in turn, that consciousness *merely accompanies* the neural processes that cause movements and that the causal relevance of consciousness is limited, at best, to a potential veto power. This interpretation suggests, in other words, that conscious P-intentions are mere epiphenomena and that the conscious self is a mere spectator in the execution of motor acts.

This is, I think, a charitable interpretation that captures the essence of Libet's challenge. It is important to note here that, on this interpretation, the challenge does not arise from the claim that conscious P-intentions are not the causal origins or uncaused causes of movements. In particular, it does not arise from the observation that conscious P-intentions, or their neural correlates, have causal antecedents. Rather, the challenge is generated by the fact that conscious P-intentions are preceded by unconscious neural processes that are *related and specific to* the ensuing movements.

This interpretation shows that the narrow strategy is not decisive against Libet's challenge, because it shows that the challenge is not based and not dependent on a particular psychological interpretation of RPs. In particular, the challenge is not based and not dependent on the claim that early-RPs correspond to unconscious *decisions*. Rather, the challenge is generated by the fact that *there is this neural process*, which is related and specific to the ensuing movement. Given this, we can also see that the challenge does not depend on the assumption that the RP stems from one single neural process or from activity in one specific brain area—we can see that the challenge is fully compatible with the hypothesis that the early and late parts of RP stem from different neural sources. To be sure, the empirical evidence does not demonstrate that conscious P-intentions are epiphenomena. But it *suggests* that movements are not consciously initiated, and the right interpretation of Libet's challenge shows that the narrow strategy fails to provide a decisive response.

Furthermore, the mentioned experiment by Haynes and colleagues gives rise to a very similar challenge that does not depend on the interpretation of RPs at all. In this experiment, subjects were not only free to choose when to perform a movement, but they

were also given a choice between two types of movement (pressing one of two buttons with either their left or right index finger). Using statistical pattern recognition techniques, Haynes and colleagues were able to decode brain activations from fMRI data that are predictive of both the selection of the action (what-decisions) and of the timing of the movement (when-decisions). Brain activity that is predictive of what-decisions was found as early as 10s before the conscious choices, and brain activity that is predictive of when-decisions was found as early as 5s before the conscious choices (see Soon et al. 2008).

This experiment raises a challenge that has the same structure as Libet's, and it radicalizes it in two respects. The uncovered temporal delay between the conscious choices and the unconscious precursors is obviously much greater, and the experiment raises a challenge for both when- and what-decisions. For now, however, the most important point is that this challenge is not based on the recordings of RPs. In fact, predictive neural activity was found in areas that are entirely distinct from the brain areas that are commonly associated with the generation of RPs.[2] This means that this experiment gives rise to a Libet-style challenge that does not in any way depend on the question of how RPs are to be interpreted. Further, it gives rise to a challenge that does also not depend on any particular psychological interpretation of the neural activity in question—the challenge stems from the fact that *there is this neural activity* which is predictive of both when- and what-decision, not from a particular psychological interpretation of that neural activity.

A decisive response? The broad strategy

We can already see, now, why the broad strategy is also less convincing than it may initially seem. This response points out that the Libet challenge concerns only relatively insignificant when-decisions that are subordinate to earlier what-decisions. The experiment by Haynes and colleagues shows, however, that a very similar challenge arises for what-decisions as well. Moreover, Haggard and Eimer (1999) had already shown that even RP recordings can give rise to a Libet-style challenge that concerns what-decisions. Haggard and Eimer adopted Libet's experimental design, but they gave

[2] It is widely believed that RPs are generated by activity in the pre-supplementary motor area and primary motor cortex (Shibasaki & Hallett 2006, for instance). In contrast, Haynes and his colleagues identified predictive neural activity in the prefrontal and parietal cortex (Soon et al. 2008).

subjects the choice between two actions (pressing a button with either their right or left index finger), and they measured in addition the onset times of the late lateralized RP. This lateralized RP predicts which finger will move (the finger contra-lateral to the hemisphere where the lateralized RP is recorded), and Haggard and Eimer found that even the onset of the lateralized RPs precedes W (on average by about 450ms).[3] The important point for us is that subjects reported in this experiment the timing of a P-decision that has both a when- and a what-component (or that is both a when- and a what-decision).

Applying the broad strategy again, one might reply that in such experiments the subjects also form D-intentions about what to do: they consciously decide at the beginning of the experiment to press a button with either their left or their right finger. But this response is now much less convincing. It is true that at the beginning of the experiment subjects form D-intentions that generate narrow constraints on what they will do. But given that their D-intentions leave a choice about *what* to do, the challenge cannot be brushed aside in the same manner as before. Initially, the point was that the brain unconsciously initiates merely a specific type of movement that the subject consciously agreed to perform at the beginning of the experiment. This point does no longer hold. Experiments that feature free what-choices suggest that the brain may also unconsciously *select* which specific action to perform before the conscious what-choice is made.

A further problem with the broad strategy is that it seems to push the problem one step backward. The response is based on the observation that the subjects in the Libet experiment make two conscious decisions. At the beginning of the experiment, they form D-intentions concerning what type of movement to perform (within a certain time frame), and just before they move they form P-intentions to move now. The evidence from the Libet experiment concerns only the latter. But why should we think that D-intentions do not also have unconscious precursors? Why should we think that the response does not simply push the problem one step back from P- to D-decisions?

[3] This relatively long temporal gap between the onset of the lateralized RP and W is partly explained by the fact that the onset times of the RPs in Haggard and Eimer's experiment were much earlier than in Libet's (on average earlier than even 2000ms before the onset of the movement). The possible reasons for this significant difference in the onset times of the RPs are discussed in Haggard and Libet 2001.

On the basis of the Libet experiment, one might argue as follows. There is no reason to think that the problem reappears for D-decisions, because there is no reason to think that D-decisions are accompanied by RPs. There is, firstly, no empirical evidence that suggests that D-intentions are accompanied by RPs. There can be no such evidence, because RPs can only be obtained in relation to behavioral events. Secondly, RPs are commonly associated with *motor preparation*. D-decisions, however, are usually made well ahead of their execution. One can, for instance, decide now to go shopping tomorrow without moving a finger, as it were, and without preparing to move. Similarly, being given the instructions at the beginning of the experiment, subjects form D-intentions to perform a certain actions later, during the experiment. Given this, it is simply implausible to think that D-decisions are accompanied by RPs, because it is implausible to think that they are accompanied by neural processes of motor preparation.

But this response is not convincing, because it addresses only one possible source of the Libet-style challenge—namely, the recording of RPs. As mentioned, Haynes and colleagues have identified unconscious precursors in brain areas that are not associated with motor preparation or RPs. This raises a Libet-style challenge that is not based on the recordings of RP. Moreover, recently Haynes and colleagues have shown that pattern recognition techniques can also be used to identify neural predictors of consumer choices from fMRI data (Tusche et al. 2010). In this experiment, subjects did not form genuine intentions, as their choices were merely hypothetical ("I would by this product"). But due to their abstract nature, such hypothetical choices have more in common with D-decisions than P-decisions.

None of this amounts to direct evidence for the claim that D-decisions have unconscious precursors. But we have seen that there is no good reason to think that the challenge cannot arise for D-decisions, and we have seen that more recent evidence on decision-making suggests that D-decisions may as well have unconscious precursors. Taken together, this shows that the broad strategy is not decisive against Libet-style challenges.[4]

[4] Some commentators have suggested an interpretation of Libet's experiment that combines the narrow and the broad strategy. It says, roughly, that the conscious event at W triggers the execution of a movement in accordance with a conditional D-intention that has been formed at the beginning of the experiment (Keller

Two conceptions of free will

The narrow and the broad strategy highlight important issues and they have some force in response to Libet's challenge. But they fall short of providing a decisive response to Libet-style challenges. Implicitly, the narrow and the broad strategy share a crucial assumption with Libet (and with everybody who acknowledges the force of Libet-style challenges). That is the assumption that free will requires *conscious initiation*: roughly, the power to initiate actions by way of forming conscious intentions. This assumption seems plausible, and so it is no surprise that many who are familiar with Libet-style experiments acknowledge readily that the evidence poses a serious threat to free will. But there are alternative conceptions of free will that do not require the power of conscious initiation. In what follows, I will propose such an alternative, and I will argue that this conception of free will is immune to any kind of Libet-style challenge.

Within the debate in analytical philosophy, there was for some time widespread agreement that free will has basically two components. It was agreed that acting with free will requires (1) the absence of constraint (coercion and interference) and (2) the ability to do otherwise. It was also widely agreed that (1) is compatible with the hypothetical truth of causal determinism, and so most of the debate was about the question of whether or not having the ability to do otherwise is compatible with determinism (setting questions concerning moral responsibility aside).[5] The dialectic has changed in the more recent debate, partly because it has become more and more common to think about free will in terms of initiation. In particular, in an influential book, Robert Kane (1998) has argued that acting with free will requires also that the agent is (3) the *origin* or *source* of the choice and action. Kane has not formulated explicit conditions on the role of consciousness in the initiation of action. Nevertheless, it seems clear that this conception of free will is congenial to Libet's. According to Kane, acting with free will requires that the causation and explanation of an action have their sources somehow *within the agent*. Likewise, according to Libet, acting free will requires that the causation of the action is

---

& Heckhausen 1990, and Zhu 2003, for instance). But this combined response fails to put the challenge to rest for the same reasons as the narrow and the broad strategy fail individually.

[5] Van Inwagen 1983 and Kane 1998 provide extensive discussions of this debate (although both from an incompatibilist's point of view).

endogenous or *self-generated* (in the sense that it must not be "in direct response to an external stimulus" (1985, p. 529)).

In opposition to this, I shall propose a return to the conception that requires only (2), the ability to do otherwise, in addition to the uncontroversial condition (1). In particular, I shall propose an account of the ability to do otherwise in terms of counterfactual conditionals. This is a traditional compatibilist proposal, according to which free will is compatible with determinism because both (1) and an account of (2) in terms of counterfactual conditionals are compatible with determinism. Two traditional candidates for an account of (2) in terms of counterfactual conditionals are, roughly, the following:

(CA_1) An agent S is able to do otherwise if and only if
S would do otherwise, if S had *chosen* to do otherwise.

(CA_2) An agent S is able to do otherwise if and only if
S would do otherwise, if S *wanted* to do otherwise.

Taken by itself, CA_1 is unsatisfactory because its antecedent refers another action (the mental act of making a choice). This raises the question it was supposed to answer: is the agent able to do otherwise? A problem with CA_2 is that it commits us to a controversial view about motivating reasons. That is, roughly, the view that all free actions are motivated by desires. It raises also difficult questions concerning the relationship between the strength of desires and the agent's choices. What if the agent has conflicting desires? Are free choices always motivated by the agent's strongest desire? And so forth.

Further, there are problems with potential interference and intervention. It might be false that S would do otherwise, because something or someone might prevent S from doing otherwise if and only if S chooses or wants to do otherwise. Yet it may be true that S is able to do otherwise.[6] This raises difficult issues that are beyond the scope of this paper, and I will have to assume here that this problem can be met by means of *ceteris paribus* clauses, which are meant to cover actual and counterfactual interference. Given

[6] Note that the consequent of the relevant conditional may be false in the *closest* possible world in which the antecedent is true, because the mechanism of potential intervention may be present in the actual world.

this, I propose, to a first approximation, the following (recall that I assume that decisions just are formations of intentions):

(CA) (1) S is able to do otherwise if and only if:
(a) *Ceteris paribus*, S would do otherwise, if S intended to do otherwise, and
(b) S is able to intend to do otherwise.
(2) S is able to intend to do otherwise if and only if:
(c) *Ceteris paribus*, S would decide to do otherwise, if S had reasons to do otherwise.

This account requires a kind of reason-responsiveness. Changes in reasons should co-vary with changes in choices and actions. As it stands, the proposal is rather vague and also too strong. First of all, the antecedent in condition (c) allows of two rather different interpretations. A difference in reasons may be a difference in *normative* reasons or it may be a difference in *reason-states* (such as beliefs, desires, and other intentions). It is all too easy to find counterexamples to general claims about reason-responsiveness, if one thinks of it as responsiveness to normative reasons. It is also not very difficult to find counterexamples to the claim that we are responsive to what we take to be normative reasons. But it would be rather difficult to find counterexamples to the thesis that we are, in general, responsive to reason-states (on some views of this kind, it would be true by definition that all intentional actions are reason-responsive in this sense).

But I think that we do not have to retreat to an account that requires only responsiveness to reason-states, because it seems that all normal subjects—all human agents who are capable of intentional agency—are responsive to a fairly large class of normative reasons. That is, it seems that there is, for all normal agents, a fairly large class of reasons, reason-states, decisions, and actions for which it is true that the agent's reason-states would co-vary with reasons and that the agent's choices and actions would co-vary with the agent's reason-states. This is how it seems to me. Further below, we will see that there is in fact good empirical evidence for this claim. For now, note only that reason-responsiveness consists, on this view, in parts or stages: responsiveness of reason-

states to normative reasons, responsiveness of decisions to reason-states, and responsiveness of actions to decisions.

In connection with this, difficult questions arise for cases in which reasons and reason-states come apart. Suppose that S deliberates about whether to A or to B and that S judges that the reasons for A outweigh the reasons for B. Various dissociations are possible here. S may desire more strongly to B. S may in fact have normative reason to B. Or S may have normative reason to do something different altogether. So, an agent's desires may be out of line with the agent's judgment, and both desires and judgments may be out of line with what there is normative reason to do.

In the light of this, I propose an amendment of the view that is largely inspired by Fischer and Ravizza's (1998) work on reason-responsiveness. Note, firstly, that it is clearly too strong to require that, in order to be reason-responsive, S must decide otherwise in the closest possible world in which there is normative reason to do otherwise. Assuming the standard view on the truth conditions for counterfactuals, this just means that it is clearly too strong to hold that S is reason-responsive only if S would decide otherwise (*ceteris paribus*), if S had normative reason to do otherwise. Secondly, it is clearly too weak to require that the agent must decide otherwise in some (at least one) close possible world in which there is normative reason to do otherwise. The reason for this is, roughly, that this conditional dependence may be true in some close possible world due to a lucky coincidence or due to some kind of erratic response that is not in any way recognizable as a response to a reason (compare Fischer & Ravizza, pp. 65-68).

It seems that the right kind of responsiveness lies somewhere in between those two extremes. Following some of the suggestions of Fischer and Ravizza, I propose, roughly, that there must be some *pattern* of close possible worlds in which S decides to do otherwise because S has normative reasons to do otherwise, and which exhibits a minimal degree of rational consistency and coherence (concerning the choice in question). For instance, if there is a close possible world in which S has normative reason not to vote for candidate A because A has taken bribes from three big companies, and in which S decides not to vote for A because S knows that A has taken bribes from three big companies, then S is reason-responsive only if S also decides not to vote for A in a nearby possible world in which S has normative reason not to vote for A because A has

taken bribes from *four* big oil companies, and in which S makes this decision because S knows that.

This amendment allows us also to accommodate cases in which the agent decides freely to act contrary to reason. Suppose that S judges correctly that there is more reason to A than to B and that S decides and acts accordingly. For all we know, S might have decided to B contrary to reason. Similarly, S might have decided to A even if both S's reason-states (including S's evaluation of the reasons) and the balance of normative reasons had been in favor of B. In other words, for all we know, S is able to decide otherwise and it might be false that S would have decided otherwise (*ceteris paribus*), if S had reason to do otherwise—condition (c) might be violated. But the possibilities that S might have decided either to A or to B contrary to reason are fully compatible with the requirement that there must be some coherent pattern of close possible worlds in which S decides to do otherwise because S has normative reasons to do otherwise.[7] Given this, I propose to amend the account as follows:

(CA) (1) S is able to do otherwise if and only if:

(a) *Ceteris paribus*, S would do otherwise, if S intended to do otherwise, and

(b) S is able to intend to do otherwise.

(2) S is able to intend to do otherwise if and only if there is some pattern of close possible worlds W such that:

(c) There is reason to do otherwise (in each world in W),

(d) *Ceteris paribus*, S decides to do otherwise, because S recognizes that there is reason to do otherwise (in each world in W), and

(e) The worlds in W exhibit a minimal degree of rational consistency and coherence with respect to the choice in question.

[7] One must, of course, distinguish between an agent's being responsive to reasons and an agent's responding to reasons. If S decides to B because of a strong desire, then S responds to a reason-state (the desire to B), but contrary to reason and contrary to another reason-state (the judgment that there is more reason to A). If that choice is not based on any reason-state, then S simply decides to act against reason. In either case, the action is intentional, given that it is caused and guided by an intention, and S has the ability to decide otherwise, given that S is reason-responsive in making the choice.

The proposal is still rough and probably in need of further refinement.[8] But it is sufficiently precise for my purposes here. For all I want to argue is that *if* we think about free will *along the lines of* CA, then free will is immune to any kind of Libet-style challenge. The proposed account of free will is intuitively plausible, and it is important to note that it is in line with a traditional philosophical conception of free will. It is, for this reason, clearly not *ad hoc*.

In the remainder of this section, I shall briefly address the question of whether there is any evidence for CA. First of all, can there be any empirical evidence for CA? Recently at a conference I heard a speaker saying that claims about the ability to do otherwise are "not scientifically testable, because they involve counterfactuals". A first point to note here is that the experimental method itself presupposes that certain counterfactuals are true. Very roughly, an experiment is sound only if the actual distribution of subjects to the different experimental and control conditions does not make a significant difference. This is just to say that a particular experiment is sound only if a different distribution of subjects to the conditions *would not have made* a significant difference. And a different distribution of subjects to the conditions would not have made a significant difference only if a great majority of the subjects would have responded to the instructions in an alternative condition in a reason-responsive manner. Of course, for any given experiment, we can never be certain that this condition is satisfied. But it can be more or less reasonable to assume that it is satisfied.

It is true, of course, that experiments cannot directly test counterfactual claims, because they can only test what actually happens. But it does not follow that empirical science cannot provide evidence *in support of* counterfactual claims. Indeed, virtually all experiments on decision-making in cognitive neuroscience and psychology provide some support for claims about reason-responsiveness. The very simple reason for this is that, in virtually all experiments, normal subjects show that they are able to follow the instructions in a reason-responsive manner. Usually, subjects voluntarily agree to follow the instructions. This commitment gives them some reason to follow the instructions, and

[8] One might think that there is a problem with manipulation cases (such as brainwashing or insertion of beliefs and desires). But this is not obvious. If all the problematic types of manipulation undermine also the agent's responsiveness to normative reasons, then there is no problem for CA—at least not obviously so. If not, then the account must be supplemented with conditions concerning the way in which the relevant reason-states (beliefs, desires, and intentions) have been acquired.

usually subjects show that their subsequent choices and actions are guided by those reasons. Moreover, most experiments support *counterfactual* claims about reason-responsiveness, because most experiments feature more than one condition (typically an experimental and a control condition). Suppose that you are a subject in a well-designed experiment with two conditions and that you are assigned, randomly, to the experimental condition. Presumably, you will follow the instructions for this condition (*ceteris paribus*), and had you been assigned to the control condition, you would act differently in accordance with the instructions for that condition (*ceteris paribus*). It seems obvious that something along those lines holds for the vast majority of subjects in the vast majority of experiments. Further, in some experiments the same subjects are assigned to the different conditions in subsequent trials. Again, no experiment can demonstrate the truth of a counterfactual. But given that normal subjects are able to perform different tasks at different times (in accordance with different instructions), and given that the ordering of tasks is usually assigned randomly, such experiments provide at least indirect support for counterfactual claims concerning reason-responsiveness.

But there is also plenty of empirical evidence that supports the assumption of a good degree of reason-responsiveness with respect to the types of reasons and actions that feature in important real-life choices, such as reasons to take physical exercise, to wear a seatbelt, to have regular health checkups, to quit smoking, and so on. Webb and Sheeran (2006) have collected and analyzed forty-nine studies of this kind, and they have found that the evidence supports the hypothesis that interventions on agents' intentions by way of giving good reasons engender changes in their intentions and actions. In particular, their meta-analysis shows that the intervention of reason-giving has a medium-to-large effect on changes in intentions, and that changes in intentions have a small-to-medium effect on changes in behavior. One might think that this result is problematic for a proponent of CA (and reason-responsiveness), because the effect size of changes in intentions is only small-to-medium. But I think that the analysis only confirms reasonable expectations concerning the efficacy of intentions. We know that intentions are not very effective when they are up against strong habits and addictions, and we think that habits and addictions can undermine our freedom of choice. The important point is that there is

a robust effect from reason-giving all the way to changes in behavior across a wide range of real-life situations.

The worst case scenario

In this section, I will outline what I take to be the worst possible Libet-style scenario. This will be instrumental to the argument of the following section, where we will see that the proposed account of free will is immune to Libet-style challenges because it is immune to a challenge from the worst possible Libet-style scenario.

Mere reflection on our agency can reveal that many of our actions are habitual or automatic, neither based on conscious reasoning nor on conscious intention. It is an open question whether or not such actions are ever performed *with free will*. At the same time, it seems clear that some choices are made consciously and on the basis of a process of deliberation in which reasons are consciously considered and weighed. Further, it seems only plausible to think that we have *genuine* free will only if we make some such conscious decisions with free will.

The worst case scenario would be if the Libet-style challenge generalized to all the relevant conscious mental events in all instances of decision-making, including all the decisions that are apparently based on conscious deliberation. In particular, the worst Libet-style scenario would be if all the relevant conscious mental events had unconscious precursors that predict and determine their occurrence. I shall call this scenario *UD*: *universal delay* of consciousness in decision-making.

It will be helpful to illustrate this possibility my means of a causal graph. Consider first a model (figure 1) that represents a common type of conscious decision-making where the agent decides to perform an action A when the circumstances C arise. In this model, R stands for the conscious consideration of reasons, D_D for the conscious D-decision to A in C, and D_P for the conscious P-decision to execute the D-intention in C.

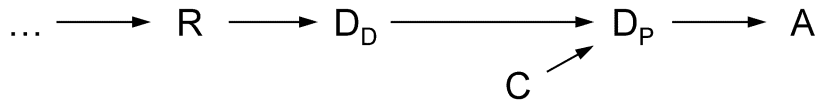

Figure 1. Causal model of conscious decision-making

In order to represent UD, I shall assume that conscious mental events are realized by neural events,[9] and I shall embed the causal graph of conscious decision-making in a model of levels of explanation (Owens 1989, for instance). In figure 2, the neural events N_2, N_4, and N_6 realize and determine the occurrence of the conscious mental events, which is indicated by the brackets. The action, as it is described at the higher level of psychological explanation, is realized by the bodily movement B. On this model, conscious mental events are realized by neural events, which have unconscious precursors (N_1, N_3, and N_5) that predict and determine their occurrence (and that determine, thereby, the occurrence of the conscious mental events).

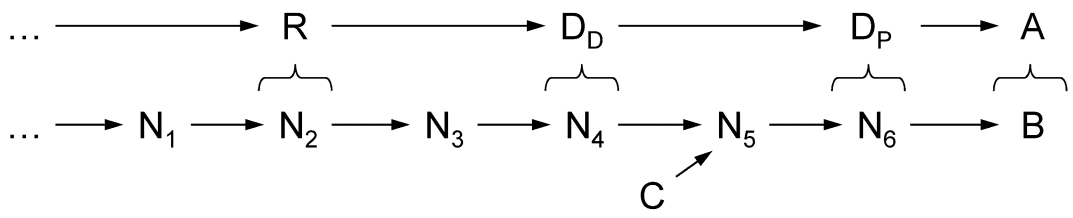

Figure 2. Causal model of decision-making with unconscious precursors

I am not aware of any experiment or study that provides direct support for UD. However, the mentioned experiments by Haynes and colleagues (Soon et al. 2008 and Tusche et al. 2010) suggest that UD is not the paranoid fantasy that it might otherwise seem to be (compare also Burns & Bechara 2007 and Falk et al. 2010). In any case, I shall assume UD only for the sake of argument. UD is the most general and most radical version of the Libet-style challenge. Given this, it seems clear that if the proposed account of free will is immune to a challenge from UD, then it is immune to any version of the Libet-style challenge.

[9] In order to accommodate externalism about mental content, one may substitute 'physical' for 'neural'. But nothing of substance hangs on this here.

Free will without conscious initiation

What matters, according to the proposed account of free will, is the co-variation of reasons, choices, and actions: if there is a change in intention, there should be a corresponding change in action, if there is a change in reasons, there should be a corresponding change in intention (in some pattern of close possible worlds). It is not difficult to see that the question of whether or not this kind of co-variation would hold is entirely independent from facts concerning the timing of conscious mental events. In particular, it may be true that conscious choices and actions would co-vary with reasons (in some pattern of close possible worlds), independently of whether or not the relevant reason-states, choices, and actions have unconscious precursors—independently of whether or not UD is true. To be more precise, the hypothetical truth of UD is compatible with the following: for any normal agent S, there is a fairly large class of actions, reason, conscious reason-states, and conscious intentions, for which it is true that S's reason-states are responsive to S's reasons and that S's intentions and actions are responsive to S's reason-states (along the lines of CA). For all that matters, on this view, is reason-responsive co-variation over sequences of decision-making, as represented in figure 3.

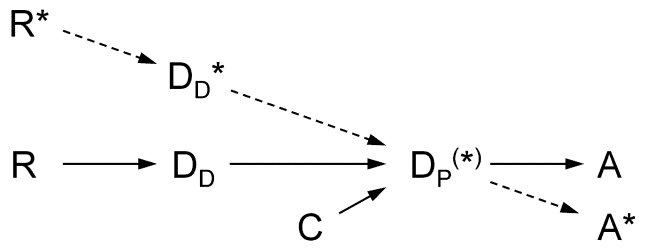

Figure 3. Causal model of reason-responsive decision-making

According to the proposed conception of free will, it does not matter whether or not actions are consciously initiated. Every conscious event in the causal sequence may have unconscious precursors. Yet it may be true that the whole sequence would co-vary in a reason-responsive manner (figure 3). This shows that this view is not subject to Libet-

style challenges at all: this account of free will is immune to any kind of Libet-style challenge, because it is immune to a challenge from UD.

Where does this leave us with questions concerning the efficacy of conscious intentions? Note, firstly, that we have already seen that there is good empirical evidence for the claim that conscious reason-states and conscious intentions are causally efficacious. The mentioned meta-analysis on the efficacy of intentions (Webb & Sheeran 2006) supports the hypothesis that interventions on intentions by means of reason-giving engender changes in intentions and actions (to some degree, at least). This supports not only counterfactual claims about potential co-variations between conscious reason-states, conscious intentions, and actions. But it supports also the claim that conscious reason-states and conscious intentions are causally efficacious in the guidance of behavior (to some degree, at least).[10]

Secondly, given that conscious intentions can be causally efficacious, it is true, *in a sense*, that we have the power to initiate actions consciously—namely, in the plain and modest sense that our conscious intentions can be among the *causes* of our actions. This does not mean, of course, that we can be the conscious initiators of our actions in the sense that our conscious intentions are the causal origins or uncaused causes of our actions. Further, it does not guarantee that we are the conscious initiators of our actions in Libet's sense, because the proposed view is compatible with the possibility that all conscious events in decision-making have unconscious precursors. (Return to figure 2 and 3. The view is compatible with the possibility that all the relevant conscious events have unconscious precursors, because the unconscious events in question may be segments of the causal sequence that realizes the reason-responsive mental causation of action.)

What's initiation anyway? What's it worth?

I have argued, so far, that if we adopt the proposed conception of free will (along the lines of CA), then free will is immune to Libet-style challenges. This conception of free will supports the claim that conscious intentions are causally efficacious, but it is

[10] This claim is based on an interventionist model of causation, which is the standard model for causal inferences in the empirical sciences.

compatible with the possibility that we lack a power to initiate that goes beyond the causal efficacy of conscious intentions. In particular, it does not require that the origins or sources of free choices lie within the agent, and it does not require that the agent must be a conscious initiator in Libet's sense. The all-important question is, then, whether or not acting with free will requires a power that goes beyond the causal efficacy of conscious intentions. It is impossible to settle this question here. But I shall briefly outline a number of considerations and arguments which suggest, jointly, that there is good and independent reason to endorse an account of free will that does not require any powers of initiation that go beyond the causal efficacy of conscious intentions—powers of origination or *genuine* initiation, as I shall call them.

Firstly, there has been plenty of disagreement, among many gifted philosophers, concerning the nature of free will. The philosophical debate has revealed, in particular, that some questions concerning the nature of free will cannot be settled by means of conceptual analysis. This includes, as far as I know, the question of whether or not acting with free will requires powers of origination or genuine initiation. The philosophical standard positions on free will are silent on questions concerning the role of consciousness for free will. But there is, in any case, no clear and widespread commitment to the claim that acting with free will requires that the agent is the origin or source of the action, and there is no clear and widespread commitment to the claim that acting with free will requires that the agent is the conscious initiator in Libet's sense.

Secondly, it is questionable that empirical research concerning commonsense intuitions on the nature of free will can settle this issue. To the contrary, empirical studies suggest that there is no such thing as one single commonsense conception of free will. In particular, pre-theoretical and commonsense intuitions can be pulled in different (and often inconsistent) directions by means of different types of examples and questions (compare Nichols 2006). It is likely that the same holds also for commonsense intuitions concerning the question of whether or not free will requires a power of initiation that goes beyond the causal efficacy of conscious intentions.

Thirdly, it is surprisingly difficult to spell out what it actually means to be the *origin*, *source*, or genuine *initiator* of choices and actions. Metaphysically, we face a dilemma. Either we assume that the agent or the conscious self can be an uncaused cause

of intentions and actions, or we assume that the agent's conscious choices can be the uncaused causes of actions. Both options are philosophically problematic (compare Double 1991, Smilansky 2000, and Pereboom 2001, for instance), and they are, as far as I can see, empirically inadequate.

A related final point is that it is also surprisingly difficult to spell out what the value of origination or genuine initiation, *taken by itself*, consists in. Free will is valuable insofar as it grounds ascriptions of moral responsibility. Arguably, free decisions ground responsibility because, and only because, they are reason-responsive (Fischer & Ravizza 1998, for instance). If we subtracted, so to speak, the agent's ability to respond to reasons from an action of which the agent is the originator or genuine initiator, then it seems that we would be left with something that is purely spontaneous and endogenous in the sense that is not made in response to anything, including the agent's own reason-states. It seems that we would be left with a *random act of indifference*—something that does not confer any recognizable kind of freedom or control. (A *locus classicus* of this line of argument is Hobart 1934. Compare also Smilansky 2000, for instance).

Conclusion

I have argued that there is an account of free will that is immune to any kind of Libet-style challenge. Taken by itself, this would show only that *if* we think about free will in this way, then we can avoid any kind of Libet-style challenge. But we can draw a stronger conclusion. The proposed account is not just *an* account of free will. It is intuitively plausible and it is in line with a traditional philosophical way of thinking about free will. Moreover, there is, as I have argued, some good and independent reason to be skeptical about the coherence and value of an alternative conception of free will that requires powers of initiation that go beyond the causal efficacy of the agent's conscious intentions. Taken together, this provides us with good reason to think about free will in the way suggested here—along the lines of the proposed compatibilist analysis of the ability to do otherwise.

There is, however, another conclusion that one could draw from the main argument presented here. I have argued, in effect, that there is a compatibilist account of the ability to respond to reasons that is immune to any Libet-style challenge. It has been argued that

the kind of control that is provided by reason-responsiveness is sufficient to ground ascriptions of moral responsibility (Fischer & Ravizza 1998, for instance). So, if one is not persuaded by the account of free will that I have proposed here, and if one thinks that reason-responsiveness is sufficient for moral responsibility, then one can conclude that the justification for ascriptions of moral responsibility is immune to Libet-style challenges, because ascriptions of the ability to respond to reasons are immune to them.

Acknowledgements

The research for this article was part of a project that is funded by the NWO (Netherlands Organization for Scientific Research). Earlier versions of the paper were presented at two workshops on free will in Amsterdam and Leiden, and I would like to thank the members of the audiences for their helpful comments. Special thanks go to David Widerker, whose comments prompted some substantial changes.

References

Burns, K., & Bechara, A. (2007). Decision making and free will: A neuroscience perspective. *Behavioral Sciences and the Law*, 25, 263–280.

Double, R. (1991). The non-reality of free will. Oxford: Oxford University Press.

Falk, E.B, Berkman, E.T., Mann, T., Harrison, B., & Lieberman, M.D. (2010). Predicting persuasion-induced behavior change from the brain. *The Journal of Neuroscience*, 30(25), 8421– 8424.

Fischer, J.M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. New York: Cambridge University Press.

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.

Haggard, P. & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*, 126, 128 – 133.

Haggard, P. & Libet, B. (2001). Conscious intention and brain activity, *Journal of Consciousness Studies*, 8, 47-63.

Hobart, R.E. (1934). Free will as involving determinism and inconceivable without it. *Mind*, 43 (169), 1-27.

Kane, R. (1998). *The significance of free will*. Oxford: Oxford University Press.

Keller, I., & Heckhausen, H. (1990). Readiness potentials preceding spontaneous motor acts: Voluntary vs. involuntary control. *Electroencephalography and Clinical Neurophysiology*, 76, 351–61.

Klein, S. (2002). Libet's timing of mental events: Commentary on the commentaries, *Consciousness and Cognition*, 11, 326–333.
Kornhuber, H.H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflügers Archiv Physiologie*, 284, 1–17.
Libet, B., Gleason, C.A., Wright, E.W., & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, 106, 623–42.
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Science*, 8, 529–66.
— (1999). Do we have free will? *Journal of Consciousness Studies*, 6, 47–57.
— (2004). *Mind time*. Cambridge, MA: Harvard University Press.
Mele, A.R. (2009). *Effective intentions*. Oxford: Oxford University Press.
— (1997). Strength of motivation and being in control: learning from Libet. *American Philosophical Quarterly*, 34, 319–32.
Nichols, S. (2006) Folk intuitions on free will, *Journal of Cognition and Culture*, 6 (1-2), 57-86.
Owens, D. (1989). Levels of explanation. *Mind*, 98, 59-79.
Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
Shibasaki, H., & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, 117, 2341–2356.
Smilansky, S. (2000). *Free will and illusion*. Oxford: Clarendon Press.
Soon, C.S., Brass, M., Heinze H.J., & Haynes J.D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11, 543 – 545.
Tusche, A., Bode, S., & Haynes, J.D. (2010). Neural responses to unattended products predict later consumer choices. *The Journal of Neuroscience*, 30 (23), 8024–8031.
Webb, T.L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132 (2), 249–268.
Van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon press.
Zhu, J. (2003). Reclaiming volition: An alternative interpretation of Libet's experiment. *Journal of Consciousness Studies*, 10, 61–77.