Alexander Batthyany

MENTAL CAUSATION AFTER LIBET AND SOON: RECLAIMING CONSCIOUS AGENCY

Introduction: Why Mental Causation is a Problem

Persons have conscious intentions and they perform intentional actions, that is, they do things that they mean to do. In addition, many believe that at least some conscious decisions are not fully predetermined by preceding events, but depend to some degree on themselves and their choices. The first statement refers to what is often termed mental or agent causation. The second refers to what is usually called free will, or freedom of choice. Although these two assumptions are intuitively held by many people, upon more careful reflection they turn out to be anything but obvious.

To begin with, the idea that we, through conscious choice, can affect physical structures, even if only our own bodies, implies that there exists a radically different kind of causation than that which is usually observed in nature. This wouldn't be a problem if this type of causation merely differed from the usual kind of causation – after all, many things differ from each other. The problem is that it differs fundamentally; and secondly, that the conscious will is, by all indications, not a typical 'thing'.

One difference between agent causation and other kinds of causation is that in agent causation, there is only one person to whom conscious causation is evident – the agent himself. Then there is the problem of the *modus operandi* of the supposed interaction between the agent and the rest of the world – nobody seems to have even a vague idea where and how this interaction ought to take place. In addition, there is the question of psychological (rather than physical) determinism – it does not rule out conscious causation, but it rules out free will. And so it goes on – there are the laws of the conservation of energy, which preclude any idea of consciousness as cause, unless one allows for constant breaches of this law, which then of course soon ceases to be a law. Later in this paper, I will discuss some of these objections. For the moment, let me note that all of the objections presented so far are based on purely theoretical, i.e. non-empirical arguments. They state that non-reductionist agency theories

are daring models, not least because within their framework the prospects of scientifically explaining conscious causation (let alone free will) are dim. Yet it is one thing to note that something is hard to explain and quite another to claim that therefore it does not exist. Still, from the materialist reductionist's point of view, everything seems to tell us that mental causation and free will cannot be reconciled with a truly scientific world-view. And so the theoretical debate could go on and on. But there is more to the agency debate than theoretical arguments.

The Empirical Case against Agency: Timing of Conscious Intention and Neuronal Activity

Most of the recent discussions of the problem of mental causation and free will have focused not on theoretical considerations, but on empirical findings. Specifically, there are two experimental studies – an older one by Benjamin Libet and his research team (Libet, Wright, and Gleason 1982; Libet, Gleason, Wright, and Pearl 1983), and a more recent replication by a research team led by John Dylan Haynes (Soon, C.S., Brass, M., Heinze, H.J., and Haynes, J.-D. [2008]) – which are said to demonstrate *empirically* that consciousness is not causally involved in our choices and actions. Both studies appear to be fairly simple and straightforward tests of our folk intuitions about agency and free will. Their underlying reasoning might be briefly summarized as follows: folk psychology tells us that our conscious intentions cause our movements. Therefore, we expect that the conscious intention to perform an action should occur before the neuronal and muscular events that initiate the successful execution of what we experience as a consciously caused movement. On the other hand, critics of folk psychology's concepts of mental causation and free will hold that our choices and intentions themselves are neurologically determined and that therefore neuronal activity causes both the experience of the intention and the movement. Accordingly neuronal activity should precede our choices and intentions.

In order to test these predictions, Libet and his colleagues (Libet et al. 1982; Libet et al. 1983) came up with a simple experimental protocol. Participants in the study were asked to make voluntary finger movements at their own pace while movement-related neuronal

activity in the supplementary motor area (SMA) was measured. During this simple task, subjects viewed a dot revolving around a clock face and were instructed to wait for one full revolution of the dot (2.56 seconds) and then to decide freely when to perform the finger movement. Afterwards the subjects reported the earliest time at which they consciously decided to move by reference to the position of the dot on the clock face. The time of the actual movement (M) was determined by means of an electromyogram connected to the subjects' wrists. Onset of preparatory neuronal activity, the so-called readiness potential (RP) was obtained by means of an EEG.

The experiment showed that the conscious intention to move (W) precedes the movement (M) by approximately 200 milliseconds and, more significantly for the questions of conscious causation and free will, that the readiness potential built up approximately 350 milliseconds *before* subjects consciously intended to flex their finger (RP -> W -> M; see figure 1).

This result raises the question: if we only become conscious of our decisions *after* the neuronal machinery for their execution has already been set in motion, how can the conscious will be a relevant causal factor in bringing about the finger movement? It seems as if it cannot be, because it only occurs after the appearance of the preparatory readiness potential. This indeed has been the conclusion of many.

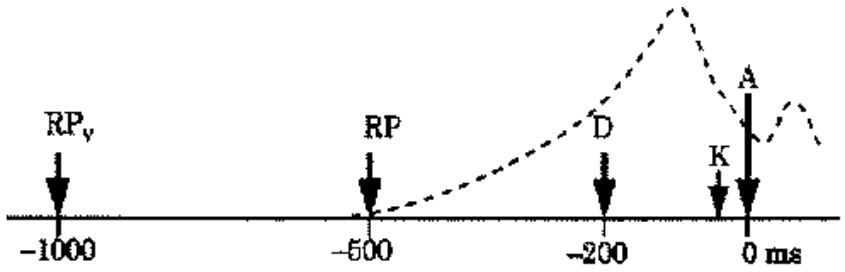

Figure 1. Simplified diagram of the results of Libet's experiments (based on Libet, 1993: 375). Sequence of the Readiness Potentials (RP), volitional decision, and onset of action (A), as well as a control stimulus on the skin (K) in Libet's experiment. If the movement is planned ahead, the RPv occurs already at –1000 msec.

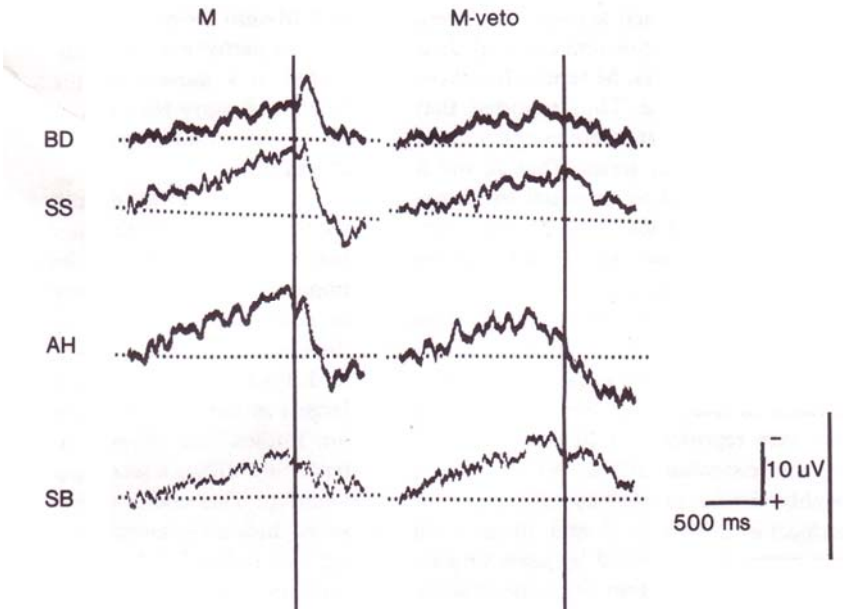

Figure 2. In the M-veto series there were no actual motor acts (as in the S series), but there was an intention to move. Subjects "vetoed" their urge. *(Reprinted from Clinical Neurophysiology (formerly Electroencephalography and Clinical Neurology), V 56, 387-372, © 1982 International Federation of Clinical Neurophysiology, with permission from International Federation of Clinical Neurophysiology).*

Nevertheless, Libet himself did not go along with this interpretation. Libet found that subjects at times felt (and reported) an urge to move (which was also preceded by a readiness potential), but for some reason decided not to yield to the movement impulse. Libet therefore assigned a veto function to consciousness. In his view, the first impulse to move is indeed not consciously caused but can be freely and consciously "vetoed" in the small window of time between RP and M (see Fig. 2). While perhaps not exactly a description of our everyday conception of free will, this model still generally holds that conscious causation and free will are real phenomena:

> It is important to recognize the almost universal experience: that we can act in certain situation with a free, independent choice and control of whether to act. […] This provides a kind of prima facie evidence that conscious mental processes can cause some brain processes. Our own experimental findings showed that conscious free will does not initiate the final "act now" process; the initiation of it occurs unconsciously. But conscious will certainly has the potentiality to control the progress and outcome of volitional processes. Thus, the experience of independent

> choice and of control (of whether and when to act) does have a potentially solid validity as not being an illusion. […] My conclusion about free will, one genuinely free in the non-determined sense, is that its existence is at least as good, if not a better, scientific option than is its denial by natural law determinist theory. (Libet 1999, 56f.)

Libet has therefore repeatedly warned his colleagues not to jump too quickly to conclusions about his experimental findings. Yet his call for philosophical caution has been largely ignored. Take, for example, the following interpretation of his results:

> Our actions are unconscious! Even though we may believe we are making a conscious decision the brain has been active half a second before that decision! The origin of our actions is not consciousness but unconscious processes. […] Consciousness leads us to believe that we can make decisions, but evidently it is no more than a ripple on the surface, or a puppet, claiming to control things that in reality it does not determine. (Norretranders 1997, 319)

Similar interpretations of Libet's findings can be found in Gazzaniga (1998, 73ff.), Velmans (2000, 211ff.), Wegner (2002), Sommerhoff (2002), Spence (1996) and Roth (1994, 2000, 2001, 2002). In an article in the popular German science journal *Gehirn und Geist*, Roth even included a photograph along with the following caption: "He exposed 'free will' as an illusion: the American neurobiologist Benjamin Libet" (Roth, 2002, 43).

It is perhaps then not surprising that the recent replication study of Soon et al. was not met with much philosophical caution either. Soon et al.'s study closely resembles Libet's original experiment, although a number of crucial modifications were introduced in the experimental protocol:

(1) the decision task was changed – instead of simply deciding when to flex a finger, subjects had to choose when to press (and which) one of two buttons with their left or right index fingers;
(2) instead of EEG measurements, a functional magnetic resonance imaging (fMRI) scanner was used, with the advantages that: (a) the activity of more brain regions could be studied (i.e. the frontopolar cortex [BA10], the medial prefrontal cortex, and the SMA and pre-SMA); and (b) this could be done for a considerably expanded time window of several seconds.

(3) Finally, instead of Libet's clock, consonants were presented in the middle of the screen, one at a time for 500 ms without gap. Subjects were asked to remember the consonant which appeared when they made their conscious decision to press a button.

Aside from these modifications in the experimental protocol, the dependent variables were the same as in Libet's experiment, with N standing for preparatory neuronal activity, W standing for the conscious decision to act, and M for the actual behavioural output. And again, as in Libet's study, it was found that the recorded neuronal activity preceded the awareness of the intention to move, though in contrast to Libet's results, activity in the frontopolar cortex and precuneus started up to *ten seconds* before the button was pressed – long before subjects reported having made any conscious decision to press one of the buttons.

Further, data from the fMRI could be used both to predict the timing of the decision and also its outcome. The researchers were able accurately to predict in 60% of trials whether the subject would press the right or left button within the next few seconds. A margin of ten percent above the mere chance expectation might appear to be relatively low, but it is statistically significant. Additionally, these predictions were made on the basis of the averaged data of 14 subjects; it is therefore possible that in-depth single-subject studies could generate an ever-higher prediction rate of decision outcomes on the basis of preceding neuronal activity.

As Soon et al.'s study has only been published relatively recently it has not been as widely discussed as Libet's experiment. Still, as with Libet's experiment, many have concluded that the results of the more recent experiment have shown that "free will takes another hammering" (BPS 2008) and the "case [is] closed for free will" (Youngstead 2008).

Given these results and their interpretation, it appears as if conscious causation, with or without free will, is an idea which these days faces perhaps the hardest test it has ever had to face. Yet I believe, and will argue in this paper, that the contemporary onslaught on agency is misguided in several ways. This paper, then, is an attempt to set right what I perceive to be a misguided debate.

An Outline of the Argument

The intention of this paper is to present a critique of reductionist interpretations of Libet's and Soon et al.'s data and to offer an alternative interpretation which is in accordance both with the empirical evidence and with the phenomenology of the will. After preliminary remarks about the causal and ontological models and the phenomenology of the will, I will argue that, contrary to claims that both experiments refute mental causation and free will, they do not in fact do so. The argument itself will be four-fold; its first aspect is mainly philosophical; whereas its second, third, and fourth aspects are concerned with the psychology of agency and its implications on the role of conscious choice and free will in these two experiments.

In the philosophical part I argue that the claim that conscious causation and free will are illusory phenomena is grounded in an ontological model – physicalism – which is bound to regard them as illusory phenomena right from the start. The trouble with the physicalist model is that it is not itself empirically verifiable; neither is it the only ontological model around. There are other ontological models – equally impossible to verify empirically – which entertain the possibility that conscious causation and free will are real phenomena. Diverse as these models are, in order to show that one cannot conclusively deny conscious causation and free will from the aforementioned experiments, it suffices to show that these models are compatible with the findings of Libet and Soon et al.; and that non-reductionist models as well as reductionist models of conscious causation and free will do in fact predict precisely the outcomes reported in these two experiments.

The second part of the argument is concerned with the psychology of agency. Here, I argue that not all willful conscious events are created equal. Some are more likely to testify to conscious causation and free will, for they seem to be dependent on an agent's act of consciously choosing and bringing them about. I will call these events *active,* or *voluntary mental events.* Others, though they have some volition-related content, are mere experiences. One key example of the latter type are desires and urges. Although these convey experiences of wanting, they are not consciously caused. For example, one cannot choose to be hungry; and neither can one choose to like certain foods, or to dislike others. Desires or urges are states or events which

are consciously experienced, but which are not subject to conscious causation and free choice. I will call such events *passive,* or *involuntary mental events*. In short, this part of the argument shows how in some cases the conscious awareness of having an intention does not represent (or depend on) a conscious decision and act of will, but is an experience – a quale of volition, whereas in other cases, the conscious awareness of having an intention does represent (and depend on) an act of will, and is not merely an experience of it.

The third part of the argument will consist of a close examination of Libet's and Soon et al.'s experiments. I will demonstrate that according to both the authors and the participants of these two studies, in the majority of the individual experimental runs, mere desire-action sequences were studied, i.e. that passive, rather than active events were the focus of these experiments. Since even strong proponents of mental causation do not claim that we can consciously bring about passive events, I conclude that there are no grounds for the claim that "the presence of unconscious pre-decisional and pre-intentional neuronal events indicates that conscious choice in general is an illusion" (Soon et al. 2008,1).

In the fourth and last part of the discussion, I will address some questions concerning introspection and illusion. I will focus on an often overlooked secondary finding in Libet's data-set indicating that the subjects had considerable insight into the nonconscious causation of their volitional experiences. Thus one of the central arguments of reductionist agency theories, the argument from illusion – i.e. the claim that we are habitually mistaken about the relation between our willings and our actions (Wegner 2002) –,will on closer analysis turn out to be inapplicable to Libet's experiment. This point applies to both Libet's and Soon et al.'s results, but since the latter does not report extensively on the subjects' introspective experiences during the experiment, the discussion will be restricted to Libet's experiment.

In this paper I do not argue that conscious causation really takes place in a critical number of cases. I claim merely that both kinds of models – both the ones that deny the reality of conscious causation and free will and the ones that affirm it – are compatible with the outcomes of the experiments. Consequently, the claim that the results of the experiments tip the scales in favour of either view cannot be justified, unless other philosophical or psychological arguments

are put forward which could provide more evidence to support one or the other view.

Ultimately, I argue that we have good grounds for believing that the experimental findings disprove conscious causation and free will only if we have independent grounds for adhering to the physicalist model. Strictly speaking, however, physicalism rules out conscious causation and free will from the outset. Hence, taking the detour to experimental findings to prove what is already presupposed would appear to be a relatively superfluous, if not circular, argumentative strategy.

Conversely, if we believe that that there is something about consciousness and the will that makes them irreducible to physical processes, the experimental findings need to be examined from this non-phyicalist perspective. Do the results contradict any of the consequences of non-reductionist theories of agency? If yes, such theories may indeed be in need of revision or should be abandoned. If not, however, the question of whether or not agent causation exists, and whether or not we have free will is as open as it was before these findings emerged. In this paper, I will attempt to show that the latter is in fact the case.

Preliminaries: Taking Philosophy and Empirical Neuroscience Seriously

Despite the fact that the majority of recent discussions of agency draw on empirical findings, the debate about conscious causation is still mainly a philosophical debate. To begin with, the questions it is concerned with are primarily philosophical: e.g. *What stuff does the world consist of? What is matter? What is freedom? What is choice? Is there room for genuine freedom, or is everything determined? What is consciousness? Is there a causal relationship between consciousness and matter? Is it unidirectional? Is it bidirectional? and so on.*

These are the questions I will address in this paper. I will show that empirical findings pertaining to these issues are indisputably relevant, but I will also argue that by no means will they enable us to answer these questions without explicitly addressing often only implicitly held philosophical assumptions. In doing so, I will attempt to avoid entering an old, and I believe relatively unproductive, debate –

namely the question as to what empirical science can tell us about consciousness and subjective states and events. One of the fundamental conflicts in the philosophy of mind is between those who believe that the brain is all that there is to the mind, and those who regard the brain as only a fragment of a more complex and rich phenomenon.

While I have some sympathy for the latter view, it is not one I adopt in the following discussion. In fact, I will have much more to say about the common ground between reductionist and non-reductionist agency theories than about what separates these theories. As outlined above, my claim is not that conscious causation and free will exist; rather, I argue that Libet's and Soon et al.'s findings do not provide conclusive evidence to the agency debate.

Reductionist and Non-Reductionist Agency Theories

Both reductionist and non-reductionist agency theories accept that some (but not all) mental events appear to depend on our consciously willing them to occur. Those that appear to be dependent on our conscious choices, I call *active* or *voluntary* events. Those events that do not appear dependent on the conscious will I call *passive* or *involuntary*. Very roughly speaking, the active aspect refers to the will, to choices, and to intentions and acts, the passive to mere experiences.

The distinction between active and passive events, however, is not simply one between overt behaviour and subjective experience. Many movements, for example, reflexes and spasms, are not consciously willed acts, even though they have a behavioural component. For example, my eye blinking can be due to the startle reflex; but it can also be an intended, active event – for instance, when I deliberately blink to give a signal to a friend. The difference between both types of events is obvious to common sense. Yet, as already pointed out, it is not at all obvious from a theoretical point of view. To begin with, the relevant causal histories of both passive and active behavioural events – eyes blinking by reflex and eyes blinking as an intentional act – are both purely physical. So according to those theories which deny ontological and causal independence to the mental, although the experience of these two types of events differ, their difference has no objective significance, let alone could it possibly indicate that the con-

scious will causes something to happen. In the following discussion, theories that follow this line of reasoning are called *reductionist agency theories*, because they reduce all appearances of conscious causation to purely physical causation.

By contrast, I understand *non-reductionist agency theories* to be theories that acknowledge not only the experiential aspect of the will, but also posit a causal aspect to conscious causation. Differ as they may in detail, non-reductionist theories of agency share an acceptance of the conscious will as the cause of at least some actions. Most of these theories accept that there are neuronal or other physical correlates of our actions, but argue that the agency involved in these actions is not reducible to neuronal or other physiological (or physical) events. Other non-reductionist theories of conscious causation go further and claim that no preceding or simultaneously occurring neuronal (or psychological, physiological, or physical) event fully determines the outcome of a conscious decision. In other words, they hold that humans have free choice.

The claims of non-reductionist agency theories are controversial. Still, it is important to note that these theories do not claim that every action is a case of conscious causation or that every decision entails free choice. Non-reductionist agency theories may, for example, hold that even though a person can consciously bring about overt behaviour, not all overt behaviour is consciously brought about; and that though a person is in principle free to choose, not all his choices are free choices.

By contrast, reductionist theories of agency are simpler than their non-reductionist counterparts. According to reductionist agency theories, for example, the basic distinction between active and passive events merely reflects a subjective reality, but does not translate into "objective causality".

Both theories therefore part ways when it comes to active events – they give different explanations of their causal histories. But they agree about the nature of passive events. To illustrate this point, and to set the background for the discussion of the findings of Libet and Soon et al., I will now turn to discuss a special case of experience which is both strongly volition-related and involuntary at the same time: urge or desire.

Boundary Conditions of Agency: The Case of Urge or Desire

Just as overt behaviour can be experienced as consciously brought about (e.g. conscious acts) or as merely happening to us (e.g. spasms or reflexes), so our experiences of the will can differ in many respects. Some can overwhelm us and convey the impression that we are not at all able to decide whether or not they will occur, although they have something volition-related about them in being experiences of the will. Desires and urges are key examples of such involuntary experiences of wanting. Further, in standard cases, desires are directed towards concrete objects; and not only is the state of desire something we cannot choose, we also cannot choose the content, or objects of our desires. Our everyday experience tells us so: if we desire something, we cannot suddenly (and successfully) decide that the desire will cease. Indeed, we often fight against an urge or desire because we believe that it is better not to succumb to that desire. In his earlier writings on free will and agency, Harry Frankfurt makes a useful distinction between first- and second-order volitions (Frankfurt 1971). The immediate desires he names first-order desires; i.e. "simple desires to do or not to do one thing or another" (ibid, 7). According to Frankfurt, a second-order desire is a desire concerned with other desires. For instance, you might have a first-order desire to eat food which you know is not healthy for you, and a second-order desire not to want to eat that food. For the purposes of this paper, when I speak of desires, I am referring to what Frankfurt calls first-order desires, i.e. desires whose existence is realized by experience alone, and not by us consciously choosing them and bringing them about. As such, these first-order desires have more in common with experiences, whereas second-order intentions are more directly related to acts of will, and are thus active events in the terminology proposed here. The important difference between active events as I define them and Frankfurt's second-order volitions however lies in the fact that Frankfurt mainly defines them to be desires about desires, whereas by active events I understand all mental events which are consciously chosen, elaborated, and intended, whether they relate to other desires or not. Still, Frankfurt's second-order volitions are useful examples for demonstrating the difference between active and passive events and their phenomenology. Probably most readers at one point in their lives have tried to resist a desire, or to control their

desires, only to find out that this requires much effort – sometimes more effort than we can muster in the moment. This is not to claim that we always fail in controlling our desires. We sometimes succeed and, according to non-reductionist agency theories, we can do so only because, in addition to being at the mercy of *passive* experiences of desire, we also have an ability to *actively* respond to the desire. Still, the way we experience some desires suggests that we cannot immediately cause them to stop even if we wish to do so. Therefore even within the framework of particularly strong versions of non-reductionist agency theories, e.g. libertarian models of free will, first-order desires are usually seen as restrictions on, and not expressions of, free choice (Kane 1996; 1999a, b). They are boundary conditions of agency, not examples of agency.

But, one may object, can we not also actively produce such desires in ourselves? We can indeed – but we can only do so by looking for (or thinking about) stimuli that bring a particular desire to the foreground of our awareness. Yet to do this we must associate an already existing weak attraction, or desire, with this real or imagined stimulus; for otherwise we would not be capable of choosing this stimulus, rather than another, as an inducement to desire. How could we otherwise know that what we are seeking is pleasing and therefore desirable? The scholastic tradition sums up this principle in the phrase: *nihil volitum quin praecognitum* – We cannot want what we have not previously recognised.

Now, in many everyday (and characteristically relatively inconsequential) situations, e.g. when selecting a dish from a menu, it is usually weak desires and inclinations that tip the balance and determine how we choose. In such cases we make what may seem to be relatively arbitrary decisions. But the arbitrariness of these decisions may be more apparent than real. In such situations we may ask ourselves what our 'gut feeling' is and then discover within ourselves a preference for one or other of the alternatives. To 'discover within ourselves a slight preference' and 'feel impelled', however, implies that these, too, are passive events: for what you discover in yourself is already there: you have not brought it about, you have merely registered it; and what you feel impelled to do is certainly not something which gives expression to your freedom to choose. Further, such initially weak desires are sometimes caused by factors to which we do not have conscious access (see Bargh & Chartrand, 1999, for an

overview of recent experimental data on this). In fact, sometimes we do not even know our desires until we ask ourselves what we desire, and even then these desires usually only transport the information content of the objects of desire and not the desires' origins (e.g. Nisbett & Wilson 1977; Johansson et al. 2005).

In sum, desires are not consciously and actively brought about – they are, therefore, passive events. Neither their occurring nor their content is something which we can consciously decide about. Non-reductionist and reductionist agency theories agree on this point for the most part, the only difference being that non-reductionist agency theories hold that humans are, at least in some circumstances, free to yield to or resist their desires and urges.

Reductionist and Non-Reductionist Predictions on the Timing of Conscious Intentions

Armed with the results of the preceding discussion, let us now turn to Libet's and Soon et al.'s studies. Proponents of reductionist and non-reductionist agency theories implicitly agree that the claim that these experiments "refute" agency has merit only if two conditions are met – though they disagree on *whether* these conditions are in fact met. The first condition (hereafter C1) is that the "freely chosen actions" in these experiments were based on what appeared to the subjects to be active events in the sense defined above. Specifically, the actions must *appear* to the performer of the action to be caused by his or her own conscious and voluntary choice. For only in the case of purportedly consciously caused and freely willed actions do the experimental results bear on the agency question. The second condition (hereafter C2) is simply that, despite the fact that the subjects really had the impression that they freely chose between different courses of action, they did not (i.e. the argument from illusion; Wegner 2002).

The reductionist's interpretation of Libet and Soon et al.'s experiments rests on the assumption that both conditions are fulfilled. And in most cases, they rest on a third, more general, thesis about what the experiment could have brought to light about conscious causation and free will. This third assumption is not necessarily shared by all proponents of agency reductionism, but it is, I suspect,

implicitly held by a majority. It states that, given physicalism, any of the possible three outcomes of the temporal relation between will (W) and neuronal activity (N) would have corroborated agency-reductionism, albeit with different degrees of certainty. Here is why. The experiment tested whether neuronal activity preceded or superseded the experience of the will in making a decision. There were three possible outcomes: (1) neuronal activity and the experience of the will's intention to move could have occurred simultaneously; (2) the experience of the intention to move could have occurred before, and presumably triggered, the neuronal activity; and (3) the neuronal activity could have occurred ahead of the experience of the intention to move. The latter is of course the actual result; but what about the other two potential outcomes? Would the interpretations differ? They would not. Indeed, according to physicalism, they cannot. The problem is that within the framework of agency reductionism, each of these three outcomes of the experiment leads to the same conclusion. Result (1) can be taken to show that the conscious will has a neuronal correlate and that the causally relevant aspect is not the conscious experience of wanting, but its neuronal correlate. Or it would have been claimed that N and W are identical, but that, given the causal and ontological supremacy of the physical, W was reducible to N. Either way there would have been no room for conscious causation and free will.

The implications of the second potential outcome, (2) prima facie seem to favor the non-reductionist theories. But instead of accepting (2) as evidence for a non-physical cause, a physicalist can interpret (2) as an indication that current measurement methods (or the ones used in the specific set-up of the experiments) fail to capture the neuronal correlates of the conscious will. In fact, Libet's interactionist veto theory has repeatedly been criticized precisely along these lines (e.g. Velmans 2002; Brass & Haggard 2007). Outcome (2) would therefore, after further refinement of the experiments, lead to outcome (1) or outcome (3). In other words, outcome (2) can be rejected as premature, and can be indefinitely dismissed with the promise that we will eventually obtain outcomes (1) or (3).

In fact, we need not wait, because (3) is the actual outcome, and we shall see whether this result supports agency reductionism. Nevertheless, it is worth noting that for the physicalist, no matter what the actual outcome of Libet's and Soon et al.'s experiments might

have been, it could have been easily taken as "empirical" evidence for agency reductionism. The common claim that the results of both experiments support physicalism is therefore perhaps not a sensible one: Physicalism is insulated against refutation precisely because it is presupposed (Batthyany 2005).

From the viewpoint of non-reductionist agency theories, however, we might ask if the two conditions C1 and C2 apply. Namely we can ask C1 did the experiment really put subjects in a situation where they believed they were exercising conscious causation and free will?; and C2 were the subjects deceived by their introspection, i.e. did they claim that the events which led to the behavioural output felt as if they were active events?

Is W an Active or a Passive Event?

Let us first address C1 as it applies to Libet's experiment. Here, the instructions to subjects were as follows:

> [...T]he subject was instructed "to let the urge appear on its own at any time without any pre-planning or concentration on when to act", i.e. to try to be "spontaneous" [...], this instruction was designed to elicit voluntary acts that were freely capricious in origin (Libet, Wright, Carlson, 1982: 324).

Subjects were instructed to *wait* until they felt an unprompted urge to move ("let the urge appear on its own"). They were furthermore explicitly asked *not* consciously to pre-plan and deliberate on when to flex their finger. Given these stipulations, it is obvious that the movement impulse is, as the instructions put it, an "urge", i.e. a passive event. Both non-reductionist and reductionist agency theories hold that such events are determined by something other than the conscious will. Seen from this angle, Libet's results merely confirm that passive events are passive events, i.e. are not consciously brought about. For spontaneous experiences of an urge are, by virtue of their spontaneity (i.e. the absence of any conscious pre-planning), passive events that appear 'on their own'; and by virtue of being passive experiences they are events in anticipation of which subjects adopted not an intentional, active mind-set but were passively waiting for the urge to occur ('without any concentration on when to act'); and that

qua desires ('urge to move, [...] capricious in origin') are events about whose origin, content and direction subjects could not consciously decide.

What about the readiness potential, then? It is obvious that the urge must come from somewhere, and both reductionist and non-reductionist agency theories agree that it is not the conscious will that brings the urge about. For the present discussion, it suffices to assert firstly that the causes of such passive events are not consciously brought about, and secondly, that as causes, they take place *before* we become aware of the urge itself since they cause the desire to emerge to begin with. The claim that Libet's experimental results empirically support reductionist agency theories is true, but it is only half true. The other half of the truth is that they also support non-reductionist agency theories' views of desires and urges. In other words, C1 does not apply to Libet's study, and consequently it cannot be claimed that Libet's results empirically refute human agency.

Let me now turn to the study by Soon et al.. The timing-related instructions in this study were almost verbatim copies of Libet's, the only difference being that subjects were additionally asked to press randomly one of two buttons instead of simply flexing a finger of their dominant hand:

> At some point, when they felt the urge to do so, they were to freely decide between one of two buttons, operated by the left and right index fingers, and press it immediately (Soon et al. 2008, 1).

Since Soon et al.'s instructions about the timing so closely resemble Libet's, it will suffice to point out again that an urge to move for the occurrence of which one has to wait is not an active, but a passive event. Hence, condition 1 is again not fulfilled.

But there exists a subtle and important difference between Libet's and Soon et al.'s instructions. Libet asked his subjects to wait for the urge to move, whereas Soon et al.'s instructions explicitly asked their subjects to act immediately on their *first* urge. In Libet's experiment, no such restriction was present, and some subjects chose to veto some of their urges to move. Soon et al., on the other hand, effectively ruled out such or similar veto-processes with their instructions. It is therefore not surprising that these researchers did not find any evidence for such a veto-function in their experimental data.

So much for the question of the timing of the subjects' movements. But what about the decision to press the right or left button? According to Soon et al., subjects freely chose which button they would press, and these decisions were preceded by unconscious neuronal processes by up to ten seconds. Even if the experimental set-up and the instructions did not give subjects any choice about *when* to move, the question remains as to whether the choice of which button to press (right or left) was an active or a passive event. The fact that the researchers were able to predict, with significant accuracy, the subjects' decision before they even became aware of their subsequent decision, makes a forceful case against mental causation. However, I will now give three arguments why the findings are not nearly as compelling as they may appear at first sight.

The first argument concerns the fact that subjects were instructed to act on their *first* impulse, or urge, to press the left or right buttons with their index fingers. Timing and motor output were thus bound to each other while subjects merely waited for the awareness that they felt that they would like to move one of their index fingers. No reasons to press this rather than that button were involved, so no active choice was asked for. Since subjects merely waited for an urge to move one of two fingers at a given time, spontaneous desire alone determined the 'decision' which button to press. And since desire is a passive, involuntary event, again both reductionist and non-reductionist agency theories predict that it is caused by something other than conscious processes.

If, on the other hand, subjects (or, for that matter, just *some* subjects in *some* of the experimental runs) did decouple timing and the left/right-decision, and indeed consciously deliberated about which button to press, but not when to do so, it remains an open question if (and when) they started consciously to plan which finger they would move when the next urge appeared. Since this possibility is not explicitly addressed in Soon et al.'s report, it is difficult to determine whether such conscious decision processes took place at all. Subjects did not report long pre-planning phases of ten seconds in length, but this in itself does not necessarily tell us much of what really happened during the experimental runs themselves. It is a well-established empirical fact that human subjects' retrospective reports of planning and deliberation processes are often unreliable, because smaller conscious units of cognitive processes are not encoded in

episodic memory and are therefore not reportable (e.g. Nisbett & Wilson 1977; Ericsson and Simon 1980).

Though the possibility of pre-planning during at least some experimental runs is speculative, there are some additional good reasons not to discard it out of hand. First of all, the situation the subjects found themselves in was fairly exceptional: subjects were asked to do nothing else but watch strings of letters on a screen while waiting for an urge to decide which button to press within the next few seconds. Readers may try this on their own – and their own experience might tell them how difficult it is *not* to think in such a situation about what they will do next, i.e. whether to press the right or the left button when the urge to move occurs the next time (see, for example, Wegner [2004], for evidence that the instruction not to think about something significantly heightens accessibility of the thoughts and intentions that are meant to be ignored).

This is not to claim that if conscious pre-planning took place, it did so in each and every trial. But it may have occurred some of the time. The fact that Soon et al.'s results were averaged measurements, and the fact that the researchers were able to predict the forthcoming motor output in only 60% of the cases might suggest that they were not dealing with one uniform decision making process, but with several, some passive, some active. Further evidence in support of this also comes from Libet's study where, despite the relatively clear and straightforward instructions, a number of subjects for one reason or another some of the time chose not to adhere to the experimental instructions and reported that they pre-planned their motor output or decided to veto an urge to act (see fig. 2 and 3).

Two tentative possibilities present themselves here. The first is that since the subjects' 'decision' when to move was determined by the first emergence of the respective urge, random fluctuations in preparatory cortical activity as opposed to conscious decision constituted the determinant of the decisional outcome of which finger was to press which button and when. In this case, again, a purely passive event took place.

The second possibility is that in some cases subjects decoupled the timing and the output decisions, with the former being determined by the urge to move (a passive event, in turn caused by cortical activity), and the latter being determined by conscious decision; in

the second case, a passive event (the urge to move) would have triggered an active event (the decision to press the right or left button).

The latter possibility deserves to be mentioned, since there are some arguments which support it. But it is speculative and clearly, there is no direct evidence in favour of it. It may well be the case that the first scenario just described is exactly what happened, at least during some of the experimental trials. But it is an overly uncertain account, and if non-reductionist agency theories were forced by the empirical evidence to make their case on this speculation alone, they would indeed be in trouble.

Let us therefore set aside for the time being the second interpretation and return to the first. This states that due to the instructional binding of the urge-determined timing and the decisional outcome, subjects had an urge not only to move, but also to move in a certain way; and accordingly – as was the case for Libet's experiment – that C1 would not apply. But here is a good objection, or rather, a good question: Is it feasible to talk about an urge or desire to press one of two buttons when in fact no desirable outcome depended on either of the buttons being pressed? An objector will say that it is not feasible; and that if not desire or urge, an active choice was the cause of the subjects' pressing the button; and that in this case the results of the experiment do indeed disprove conscious choice and free will, since the latter was shown to be preceded, and presumably caused, by non-conscious preparatory neuronal activity.

This objection, though it seems initially powerful, is, I believe, misguided. There are two counter-arguments to it. The first is less central, but for the sake of completeness deserves to be mentioned: namely that there *was* a desirable outcome to the button press – the cessation of the experimental run and thus the satisfaction of knowing that the experiment would be successfully completed (given the fact that these were volunteer subjects, there is no reason to suspect that they were motivated by anything other than the wish to be good subjects who followed instructions faithfully). Subjects had only two alternatives at their disposal to achieve this desirable outcome: press the left button, or press the right button.

Secondly, and more importantly, the objection fails to do justice to the nature of desire. It would only hold as long as one viewed urges or desires as states denoting a general, non-specific experience of wanting; for only then does the question arise whether it is feasible

to talk about an urge to press a left or a right button. However, as I have argued earlier, desires are not generic experiences, but are very concrete wants – to desire is to desire something and not to desire as such. You either desire something, or you do not desire at all. Now it is true that the experimental situation is awkward and some readers might find it difficult to imagine ever having a desire to press a button with either their left or right index finger. But given that subjects had only these two alternatives, their possibilities to experience and fulfil any desires were severely restricted. Arguably, it is this extreme restriction which is difficult to imagine, not the fact that under highly restricted circumstances only few – in this case, only two, and somewhat unusual – urges will come to mind and be acted out. What for the moment counts, therefore, is not so much that such a situation seems to be so remote from our day-to-day lives. Rather, what matters is that whatever the specific content of the urges to move this or that finger, the button presses were based on these urges rather than on conscious deliberation or choice, in which case we are again talking about passive events.

To summarize, there are several good grounds to question whether C1 was fulfilled in Soon et al.'s experiment. While I treated them as if these were separate arguments, it is possible that more than one of the mechanisms I have mentioned were simultaneously operational during the experimental tasks. When it comes to the question of whether Soon et al.'s results provide evidence in support of non-reductionist and reductionist agency theories, it matters little whether all subjects experienced a concrete urge to move the right or the left finger on every occasion, or whether some subjects at times simply wished to terminate a trial by randomly pressing one of the two buttons, or whether timing and the urge to move were in other ways bound to each other.

In all three cases something other than an active event stood at the beginning of the action sequence, and thus, in all three cases non-reductionist agency theories would predict exactly the same outcome as reductionist theories.

I conclude that Soon et al.'s fMRI study on binary behavioural choice provides no more relevant empirical evidence pertaining to the question of conscious causation and free will than does Libet's original study. Both shed light on the causal history of an urge, or desire, to move; but in both, C1 was not fulfilled. Since C1 is not met,

the findings themselves do not specifically support either view on mental causation and free will.

Introspection and Illusion

I will now proceed to discuss the second condition, C2 – the argument from illusion. But before doing so, one has to keep in mind that C2 is partially dependent on C1. For only if C1 holds – namely that subjects believed they were actively choosing their movements – can it be said that subjects succumbed to the illusion of having consciously chosen and caused their movements when in fact they did not. In the preceding section I attempted to show that C1 does not apply to either of the experiments, so one may wonder whether C2 is applicable at all. It is not, at least not in the way set forth in the introduction. But by making a slight shift in emphasis, C2 can still be used to construct an interesting test for non-reductionist agency theories. Because in the experiments no genuine conscious causation was stimulated, the question of how the subjects experienced their passive events gains in importance. For instance, one might expect subjects to have at least partial insight into the purely experiential, passive nature of their urges to move. After all, non-reductionist agency theories usually claim that there is some truth value in introspective reports. Reductionist agency models, on the other hand, suggest only a very loose relationship between experiences and supposed acts of will, thus making it more likely that here too subjects would make mistakes in their introspective reports. It is entirely consistent with reductionist theories of agency to deny validity to introspection, though note that this strategy also undermines the force of the very findings that are claimed to underpin the argument from illusion (C2). After all, the philosophical significance of the data of both experiments rests to a considerable degree on the apparent discrepancy between subjective experience and the actual measurements of the temporal onset of movement-related neuronal events. There is therefore an awkward circularity in the appliance of the argument from illusion to the experiments. On the one hand, it is presupposed that first-person accounts are to a large extent non-veridical depictions of what is "really" happening in brains and minds; on the other hand, in Libet's and Soon et al.'s studies, first-person accounts are one of the

primary means used to "prove" that first-person accounts are unfit to be the basis of a theory of agency. I do not think that this selective usage of first-person data is an irreparable defect of reductionist agency theories, but nonetheless it is worth pointing out that this is not a very straightforward way to handle data and its interpretation.

The crucial question is: What did the subjects themselves report about their subjective experiences during the experiment? It turns out that subjects evidently had precisely the kind of insight that is called into question by the proponents of C2. When asked how they experienced the task, the subjects reported that

> […] each urge or wish to act appeared suddenly out of nowhere, with no specific pre-planning or pre-awareness that it was about to happen (Libet *et al.*, 1983, 638).

One should not expect untrained subjects to have the sort of informed conceptual understanding of neuronal processes that would equip them to express their "out of nowhere" in the language of the behavioural scientist or the philosopher of mind. It is quite sufficient that they correctly state that their movement impulses were not the object of conscious reflection, intention and willing ('no specific pre-planning'). So, contrary to C2, subjects did not claim that they consciously decided their movements.

Another only rarely acknowledged secondary finding in Libet's data-set strengthens the argument for the introspective reliability of his untrained subjects: At irregular intervals they reported different kinds of W (pre-planning, the urge to act, spontaneity of a kind that surprised them, and veto) and these types of will were seen to significantly co-vary with changes in readiness potentials (Libet *et al.*, 1983, 636, see Fig. 3):

> For some series of trials the subjects reported having pre-planned a range of clock time in which they would act, in spite of our encouragement not to do that. Those series produced RPs (#1) with earlier onsets, averaging about -800 to -1.000 msec (before the motor act). […]
> In those series of forty acts in which the subjects reported no pre-planning of when to act, the onset of the RPs (#II) averaged -550 msec (before activation of the muscle). (Libet 2004, 130).

These results provide us with additional evidence suggestive of subjects' ability fairly accurately to self-report and represent at the same time an additional reason why C2 is not met in Libet's findings. Hence neither C1 nor C2 sit well with Libet's original data: C1 is simply not fulfilled; and C2, even the revised form, is if anything, supportive of non-reductive agency theories in that they predict that subject's introspective reports can be considered to be reliable data themselves.

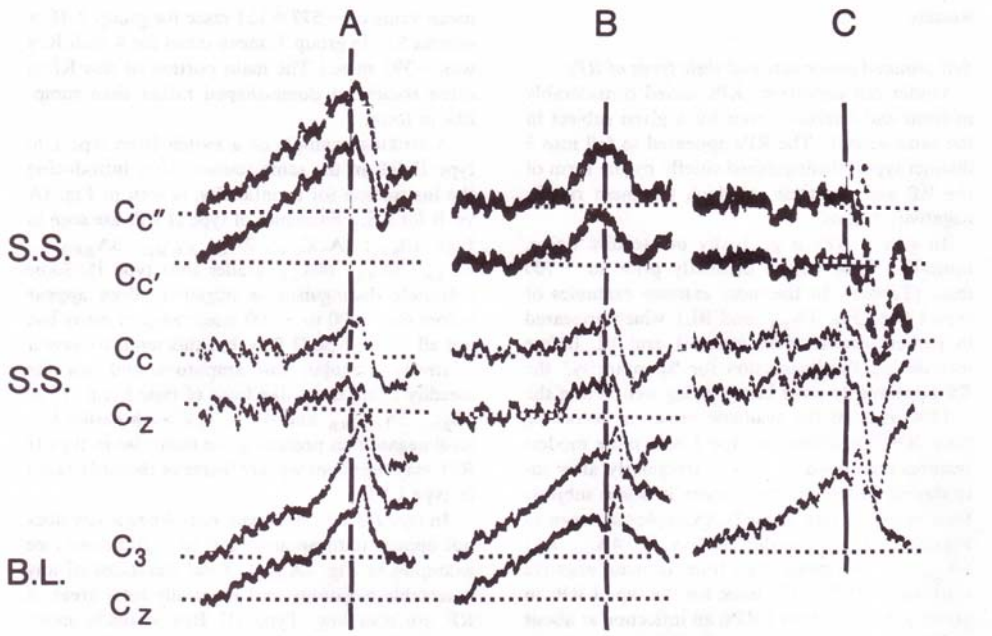

Figure 3. Self-initiated readiness potentials with different instructions and subjective reports. Each horizontal row represents 40 average potentials. Readiness potentials A were followed by the subject's subjective reports of pre-planning the movement; readiness potentials B were recorded after the explicit instruction to "let the urge … to act" spontaneously. *(Reprinted from Clinical Neurophysiology (formerly Encephalography and Clinical Neurology), V 54, 322-335, © 1983 International Federation of Clinical Neurophysiology, with permission from the International Federation of Clinical Neurophysiology).*

It is certainly true that human subjects sometimes err in ascribing to themselves conscious causation of an event. But the point here is not that humans sometimes commit such errors. Rather, it is that according to reductionist agency theories, they always do so when it comes to supposedly active events. Yet in Libet's experiment, it seems that subjects were well aware of the fact that they did not con-

sciously bring about their urge to move. So, whatever can be said about C2's status, it certainly does not represent a general and universal fact about human introspection and agency. According to Libet's results the subjects prove to be accurate in spite of the fact that they are reporting about unconscious events. Thus, C2 does not apply either – neither generally, nor to Libet's experiment.

Conclusion

Contrary to the reductionist interpretations of the findings of Libet and Soon et al., it is no objection to conscious causation that it does not entail causing urges or desires. For urges or desires are passive experiences rather than actively and consciously chosen mental events; both empirical psychology and our everyday experience tell us that much, and so do Libet's subjects when they report that they did not consciously bring about their urges to move, but that the urges came "out of nowhere". Importantly, non-reductionist agency theories, too, predict that desires and urges are not consciously chosen and brought about. I therefore conclude that neither Libet's original experiment, nor the follow-up study by Soon et al. can be legitimately interpreted to provide empirical evidence in favour of agency reductionism.

More generally, the lesson we can draw is that it is highly problematic to study conscious causation in cases where the subjects themselves state that they did not consciously cause the act in question. One cannot, for example, passively wait for an urge to occur while at the same time being the one who is consciously bringing it about; and there are many similar situations where a blatant disregard for first person phenomenology leads one to philosophical interpretations of empirical data which support one's philosophical bias only because such a bias guides one's reasoning in the first place. Yet there are logical limits and psychological boundaries to what events can be consciously caused. If one attempts to look for conscious causation and free will beyond these limits and boundaries, it should come as no surprise if neither are found.

Acknowledgements

I wish to thank Dimitri Constant for his helpful comments and criticisms of earlier drafts of this paper. Grateful acknowledgement for proofreading and correcting go to Gordon C. Wells, Benedict Coleman, and Plum Webber.

References

Bargh, J.A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462-479.

Batthyany, A. (2005). Neurophilosophie des Spontanverhaltens. In: Peschl, M. (2005). *Auf der Suche nach dem Konzept/Substrat der Seele. Ein Versuch aus der Perspektive der Cognitive (Neuro-) Science.* Würzburg: Königshausen und Neumann

Bolbecker, A. R., Cheng, Z., Felsten, G., Kong, K-L., Lim, C. C. M., Nisly-Nagele, S. J., Wang-Bennett, L. T., & Wasserman, G. S. (2002). Two asymmetries governing neural and mental timing. *Consciousness and Cognition*, 11, 265–272.

British Psychological Society Research Digest Blog (2008, April). Libet redux: Free will takes another hammering. Retrieved April 2008 from http://bps-research-digest.blogspot.com/2008/04/libet-redux-free-will-takes-another.html

Breitmeyer, B. (1985). Problems with the psychophysics of intention. *Behavioural and Brain Sciences* 8, 539-540

Bridgeman, B. (1985). Free will and the functions of consciousness. *Behavioural and Brain Sciences* 8, 540

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of social information processing goals: Nonconscious priming reproduces effects of explicit conscious instructions. *Journal of Personality and Social Psychology*, 71, 464-478.

Frankfurt, H.G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, Vol. 68, No. 1:5-20.

Gazzaniga, M.S. (1998). *The mind's past*. Berkeley: University of California Press.

Johansson, P., Hall, L., Sikstrom, S. & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310, 116-119.

Kane, R. (1996). *The significance of free will.* New York: Oxford University Press.

Kane, R. (2000). The dual regress of free will and the role of alternative possibilities. *Philosophical Perspectives*, 14: 57–79

Kane, Robert. (1999a). On free will, responsibility and indeterminism. *Philosophical Explorations*, 2: 105–21.

Kane, Robert. (1999b). Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy*, 96: 217–40

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioural and Brain Sciences* 8, 529-539

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, 6, No. 8–9: 47–57

Libet, B. (2003). Can conscious experience affect brain activity? *Journal of Consciousness Studies*. 10: 12, 24-28(5)

Libet, B., Gleason, C. A., Wright, E. W., and 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*, 106:623-642.

Libet, B.; Wright, E. W.; Gleason, C. A. (1982). Readiness potentials preceding unrestricted spontaneous pre-planned voluntary acts. *Electroencephalographic and Clinical Neurophysiology* 54: 322–325.

Miller, J. and J. A. Trevena (2002). Cortical movement preparation and conscious decisions: averaging artefacts and timing biases. *Consciousness and Cognition* 11, 308-313.

Nelson, R.J. (1985). Libet's dualism. *Behavioural and Brain Sciences*, 8, 550

Nisbett, R., T. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.

Norretranders, T. (1997). *Spüre die Welt. Die Wissenschaft des Bewußtseins*. Reinbek: Rowohlt

Roth, G. (1994). *Das Gehirn und seine Wirklichkeit. Kognitive Neurobiologie und ihre philosophischen Konsequenzen*. Frankfurt/Main: Suhrkamp

Roth, G. (2000). *Die Pseudoherrschaft des Ich*. Interview with Arnulf Marzluf http://www.suhrkamp.de/buecher/roth/rothlinks.htm

Roth, G. (2001). *Fühlen Denken Handeln: Wie das Gehirn unser Verhalten steuert*. Frankfurt/Main: Suhrkamp

Roth, G. (2002). Gleichtakt im Neuronennetz. *Gehirn und Geist* 1, 38-48

Sommerhoff, G. (2002). *Understanding consciousness. Its function and brain processes*. Sage Publications

Soon, C.S., Brass, M., Heinze, H.J., and Haynes, J.-D. (2008a). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11: 543.

Soon, C.S., Brass, M., Heinze, H.J., and Haynes, J.-D. (2008b). Unconscious determinants of free decisions in the human brain. *Suppl. Information. Nature Neuroscience* 11, Suppl.

Spence, S. (1996). Free will in the light of neuropsychiatry. *Philosophy, Psychiatry and Psychology* 3, 75-100

Velmans, M. (2000). *Understanding consciousness*. London: Routledge

Wegner, D.M. (2002). *The illusion of conscious will*. Boston, MA: MIT Press

Youngsteadt, E. (2008). *Case closed for free will*. ScienceNow Daily News 14 April 2008

In:

Batthyany, Alexander & Elitzur, Avshalom C. (2009)

Irreducibly Conscious.
Selected Papers on Consciousness.

Heidelberg: Universitätsverlag Winter