Sveinung Sundfør Sivertsen

Do or don't

Why neuroscience hasn't settled the question of free will

(and a hint at a different answer)

Master Thesis in Philosophy

Spring 2013

Supervised by *Jonathan Knowles*

Faculty of Humanities

Department of Philosophy

NTNU Trondheim

Cover design and image by Sveinung Sundfør Sivertsen, © 2012-

Cover and title pages set in Adobe Myriad Pro and Adobe Minion Pro

Main body text set in Adobe Caslon Pro 12 pt.

Printed by NTNU Trykk, Trondheim 2013

Abstract

In recent years, scientists and science popularisers alike have seen profound consequences for our view of ourselves and the organisation of society in new findings about the functioning of the human brain. Prominent in the debate surrounding these claims is the question of free will, i.e. whether or not humans are able to choose and act freely in a certain fundamental sense thought required for our practice of holding ourselves free and responsible for our actions, both morally and legally. One common position, as taken by, e.g. Sam Harris (populariser) and Daniel Wegner (scientist), holds that free will of this kind is unsupportable in the face of empirical evidence – i.a. evidence from neuroscience about the way consciousness lags behind unconscious neural processes – and that we therefore need to revise our views and practices in light of these scientific facts.

In this thesis, I argue that what might be termed the "revisionist" position is predicated not only on empirical evidence, but is essentially motivated by a belief in the fundamental incompatibility of free will with any reasonable (meta-) physics. In Part 1 I investigate the fundamental philosophical debate and find that the question of the possibility of free will is unresolved, thus challenging any simple appeal to the impossibility of free will such as that made by Harris in his short book on the subject, *Free Will* (2012). I also provide independent reason for upholding a broadly commonsense belief in free will by highlighting the sceptical nature of the challenge from determinism, which can be overcome with the help of P.F. Strawson's "soft naturalism"-appeal to our self-justified reactive attitudes. In Part 2 I investigate the empirical evidence adduced as support for the revisionist position, focused through the well-developed argument presented by Wegner in his *Illusion of Conscious Will* (2002). Here I argue that the revisionist interpretation of the data loses out to a traditional interpretation that is realist about conscious causal efficacy when the former is divested of its untenable appeal to incompatibilism.

I conclude that neuroscience has not settled the question of free will, and, furthermore, that the current state of the two debates – the theoretical and the empirical – supports a continued belief in free will of a kind that fits with our practice of generally believing ourselves free in our choices, and responsible for our actions.

Contents

1 Introduction

Reading some of the more popular accounts of research on the evolutionary history, psychological mechanisms, and physical substrates of the mind, one can easily get the impression that the results are mostly bad news for our traditional conception of our thoughts, our emotions, our actions and ourselves. Scientists-authors such as E.O. Wilson, Richard Dawkins, Robert Wright, Sam Harris, Jonathan Haidt and Daniel Wegner exemplify what might be considered a scientifically founded call to fundamental reform – be it of our mind-set, the judicial system, or society in general.

Limiting the scope to neuroscience-informed critiques of common sense and traditional ideas, two of the more prominent voices in recent years belong to the popular science writer (and polemic) Sam Harris and the psychologist Daniel Wegner.

The puppet master cover for Sam Harris' book *Free Will* (2012), opens up to a short but definite rejection of all talk about anything like free will, the ability to do otherwise, and any notion of responsibility that depends on these, importantly inspired by experiments like those of Benjamin Libet and others that appear to show that conscious choice comes too late to be the cause of action. This, purportedly, ends the age-old debate about the paradoxical facts that seem to underlie our existence as choosing individuals in a determined world, namely the apparent impossibility of any kind of "free will" in a world where every event is determined by the unyielding laws of physics and events long in the past.

Not quite satisfied with the simple rejection of free will as obviously impossible (the idea of free will has historically proved itself recalcitrant to such attacks), Daniel Wegner sets out for a long haul in his book *The Illusion of Conscious Will* (2002), drawing on myriad anecdotal and empirical evidence to show why this feeling of being in control is so hard to shake. With a drawing of a mechanical doll on its cover, *ICW* argues that our experience of willing things or being free is nothing more than that, an experience, and does not tell us the truth about what is going on: The experience of conscious will is an illusion generated alongside our actions by whatever neural mechanisms are the true causal springs of behaviour.

❁

On the one hand, you could consider this business as usual. Science is appearing in its accustomed role as purveyor of objective fact, exposing the faults and fallacies of pre-scientific speculation for the good of knowledge and the advancement of humanity – this time around, taking on what we

thought we knew about ourselves as acting agents with freedom of choice and responsibility to match. On the other hand, few other topics of research engenders such controversy as this, and for good reason: neuroscience and its cognates are here butting up against a vast amount of everyday experience and common sense, often elucidated in exceedingly intricate philosophical debates going back thousands of years.

Both Harris' and Wegner's books are concerned with roughly the same topic, namely the apparently poor fit between existing popular ideas about ethics, especially in terms of choice and responsibility, and what science seems to show us about the limited role of consciousness and the physical nature of our brains. This reminds us that it is not *Homo sapiens* as such that makes for controversial science (I doubt anyone would be offended by the elucidation of the precise mechanisms of cancer metastasis), but rather studies of the roots of our *meaningful experience*, be it consciousness, rationality, emotions, morality or value. Perhaps one could say that the friction comes from objective science stepping onto the home turf of subjectivity. A doctor explaining that the pain in my stomach is due to a bacterial infection is, from my perspective, doing something quite different from a neuroscientist telling me that the same pain is just activation of the parieto-insular and anterior cingulate cortices of my brain; even though from a scientific point of view the one could be seen (merely) as an extension of the other. Thus, experiments using advanced imaging techniques, ever more precise knowledge about the functional anatomy of the central nervous system, and the ability to elicit or modify behaviour by psychological and/or direct physical manipulations that can bypass conscious control, have all stirred up significant debate about the veracity of subjective experience itself. These books claim that we are importantly wrong about things that up until now have seemed perfectly and obviously true, and they do so with the help of, i.a. neuroscience.

Concomitant with making the human mind a subject of scientific study there is also a drive to describe the phenomena studied in terms compatible with existing scientific language, an endeavour which in this case often amounts to "naturalising folk psychology", i.e. redefining the terms in which humans usually talk of other humans' thoughts and actions in terms more amenable to the kind of precision demanded by science. Failing such redefinition, there has been a tendency to claim that the original terms do not track reality as described by science, and, giving preference to science, to conclude that folk psychology presents a false image of the world. This is what has engendered most controversy in the current case, since the terms proposed to be redefined or thrown out by advancing science are ones that people care a great deal about, and sometimes terms on which rests much established ethical theory and legal practice.

Radical as it may be, proponents of a science-based revision of ethical concepts and practice usually hold this to be a change for the better, aiming as it does to increase the precision of ethical discourse (and by extension, legal) by making sure that one only utilizes terms that have a grounding in natural fact. Just as we should use our most precise mathematical equations and knowledge of physics when launching astronauts into space, our moral language should be as precise and well-grounded as possible to avoid mistakes and bad decisions. If it turns out that concepts like "free will" and "conscious control" are poor guides to understanding and judging human action, we should stop using them altogether, substituting instead terms that track real phenomena in the physical world. As already noted, this aim of science and its interpreters not only puts them at odds with commonsense ideas and folk psychological terms – there is also the matter of ethical theory, the philosophy of which has historical roots as far back as written history permits us to look.

While no matter on which new generations will bother to pronounce can be immunized against sweeping statements by the mere existence of previous debate, there seems to be a particular willingness on the part of some scientists to say controversial things about human experience based rather simple laboratory experiments. However, many of the peer-reviewed articles on which the above mentioned popularisations are based are far more moderate in the claims they make for the import of their results, so much so that the most striking claims of these books seem more like hyperbole than necessary conclusions to an argument. Furthermore, being a debate, there are strong dissenting voices, even if they are not as widely read or easily spotted. There are both scientists and philosophers who disagree with conclusions like those mentioned above, mainly either by attacking the argumentation itself as more or less unsupported by the evidence adduced, or by granting the truth of the conclusions in principle, but denying that it has the kind of consequences for traditional thinking that the polemics claim it has.

In this thesis, I will approach the matter of whether neuroscience has settled the question of free will by a two-step analysis, dealing first with the fundamental philosophical debate concerning free will and determinism – a debate both Wegner and Harris appear to regard as settled to the disadvantage of free will – before moving on to evaluating the empirical evidence brought forth to argue against free will, centred on Wegner's argument about the illusory nature of our experience of conscious will. Following the lines of the analysis, my argument will also be two-fold, establishing first a theoretical basis for the possibility of free will, before moving on to argue that the empirical evidence, by supporting the causal efficacy of consciousness, likewise can be taken to bear out a substantial notion of free will.

Part 1

Neuroscience and the philosophical debate concerning free will and determinism

2 Neuroscience, Free Will and Determinism

> "All theory is against free will; all experience for it." Samuel Johnson (1709-1784)[1]

> "If the scientific community were to declare free will an illusion, it would precipitate a culture war far more belligerent than the one that has been waged on the subject of evolution" Sam Harris (2012, p. 1)[2]

2.1 The Limits of Neuroscience

As noted in the introduction, the popular science literature is awash with bold claims about how new science is exposing the falsehoods and illusions to which, supposedly, our traditional thinking about morality and freedom is prey. However, at the heart many of these "challenges to common sense" lie philosophical conundrums that have never been, and arguably never can be, subject to the kind of empirical testing which is the province of scientific research. Identifying these foundational issues and their place in the contemporary, science-informed "revisionist" project is an important step in the evaluation of these challenges.

One of the philosophical conundrums forming the background for an important part of the literature on neuroscience and morality is the difficulty of fitting human freedom into a coherent metaphysical worldview. The discussion surrounding this is complex and known under different names, but I will be referring to it here as "the debate concerning free will and determinism".

What is this debate – why is there a problem with fitting free will into our view of the world? It is today widely accepted among both philosophers and scientists that the world and all its inhabitants exist within the same, unified "framework"; that we are all part of the same nature, to put it prosaically. The trouble with this is that what we know of the mechanisms of physical nature appears to leave no room for the kind of freedom so obvious in our everyday experience as acting subjects, as "agents". Nature's building blocks are governed by laws, and unless something as yet inexplicable happens when those blocks build up humans, we too are ultimately governed by the same laws. If the world is *deterministic*, these laws govern change in such a way that two identical starting positions will always develop in the same manner. In virtue of this, they seem to undermine the freedom we believe to be in possession of when we think that we could have done otherwise in

[1](Boswell [1791] 2012).

[2] Also, a note on footnotes: There will be some of them throughout the thesis, but they are never essential to understanding the main argument, and can therefore be ignored. I have tried to keep their numbers to a minimum, but in places where I anticipate objections that are only somewhat related to the argument, and in places where I wish to point out interesting or important aspects of what I am discussing that is only tangentially relevant, I have allowed myself the luxury of writing compact little notes with questionable clarity to point the reader in various directions of interest.

the past, and that our choices in the present are real; that the future is yet to be decided. For, given that the world was a certain way at some time in the distant past, there appears to be but a single way for the world to be today, so also for tomorrow and all foreseeable future. It appears that this is a world where everything is laid out already. If, on the other hand, the world is *indeterministic* – i.e. *not* deterministic – it is either governed by laws that are *probabilistic*, or is also partly or completely *lawless*. But probabilistic laws appear to be no more amenable to our freedom than fully deterministic ones, in that they only introduce an element of chance to the proceedings over which it is difficult to see that we could have any control. Indeed, some argue that probabilistic laws would make us *less* free, since the only difference from strict determinism would be the chance that our choices would sometimes fail to reflect the deliberation preceding them, or our actions fail to reflect our choices. A partly or wholly lawless universe is an intriguing possibility, but advocates of such a solution will find similar difficulty in explaining the connection between the reasons we have for acting as we do and the unprecedented, uncaused, undetermined actions themselves (in this way it could be said to present a conceptual threat to rationality). Whatever your attitude to the question: theory does indeed seem to contradict experience in this case.

While Johnson's exasperated exclamation perfectly captures the paradox that by then had already troubled thinkers for millennia, Sam Harris appears to be suggesting that now, some 270 years later; science can finally decide the issue.

2.1.1 Harris' Trident

Is science in a position to decide the question of free will? That depends. It depends first and foremost on what is meant by "free will", and this in turn depends on the overall project of the person setting out to answer the question. In his short book on the subject, *Free Will* (2012), Sam Harris starts with the following:

> The popular conception of free will seems to rest on two assumptions: (1) that each of us could have behaved differently than we did in the past, and (2) that we are the conscious source of most of our thoughts and actions in the present. As we are about to see, however, both of these assumptions are false. (2012, p. 6)

The two assumptions appear to be independent, with (1) seeming like a classic target for determinism, and (2), something to which neuroscience might speak. Harris, however, effectively treats them as a single unit, and argues from three sources of support to the conclusion that free will is nonsense. He takes "free will" to be a) incompatible with both determinism and indeterminism, b) refuted by neuroscience, and c) not even supported by our own subjective experience (Harris, pp.

5-6). Because the three strains of his argument are so tightly interwoven, it is difficult to say when he is appealing to what. Indeed, I think it reasonable to say that he regularly appeals to all three, with each part-claim simultaneously giving support and being supported by the others. Thus, in the opening chapter of his book, the three elements (a), (b), and (c) come together as Harris writes:

> Free will *is* an illusion. Our wills are simply not of our own making. Thoughts and intentions emerge from background causes of which we are unaware and over which we exert no conscious control. (Harris 2012, p. 5, emphasis in original) (b)

> Free will is actually more than an illusion (or less), in that it cannot be made conceptually coherent. Either our wills are determined by prior causes and we are not responsible for them, or they are the product of chance and we are not responsible for them. (2012, p. 5) (a)

> But the deeper truth is that free will doesn't even correspond to any subjective fact about us–and introspection soon proves as hostile to the idea as the laws of physics are. Seeming acts of volition merely arise spontaneously (whether caused, uncaused, or probabilistically inclined, it makes no difference) and cannot be traced to a point of origin in our conscious minds. A moment or two of serious self-scrutiny, and you might observe that you no more decide the next thought you think than the next thought I write. (2012, p. 6) (c)

The problem with this way of combining seemingly hedged conclusions is that it obscures the fact that Harris' entire conception of the issue is constrained by his holding all three claims to be undeniably true. When Harris in c) claims that introspection alone – what has usually served as the strongest ally of the idea of free will – can reveal to us the falseness of that belief, he *appears* to be stating a self-evident fact, but is actually presenting a zero-option scenario defined by his implicit and unfounded claim that free will necessarily entails being something like the conscious ultimate cause of oneself, *and* that this is an impossible requirement. It might very well be an impossible requirement, but Harris presents no real argument for this, nor for why we should think free will would require something like being the conscious cause of oneself. Which is not to say that no such argument is possible; Galen Strawson[3] is responsible for one of the best-known versions of the argument that free will is impossible because it would entail you having to be the cause of yourself ("*causa sui*"), an entailment that arguably leads to a vicious infinite regress of "you" causing "you" (the "Basic Argument", Strawson 2010). Indeed, Harris implicitly acknowledges the source of this argument by thanking Strawson for his input, but while Strawson's argument is influential, it is by

[3] Famous philosopher in his own right, Galen Strawson is also the son of P.F. Strawson, to whom we will turn in Chapter 5.

no means universally accepted, and so Harris is making a substantial assumption that he neither properly acknowledges nor defends in relation to the wider field of the debate concerning free will.

Harris' claim about introspection is insufficiently substantiated to stand on its own, and can therefore arguably only be understood in relation to his position on the two other issues, namely the claim that neuroscience has shown consciousness to lag behind the actual causes of thoughts and behaviour – causes which he thinks are to be found in non-conscious neural processes –, and the theoretical position that free will is incompatible with what we know of the world being deterministic or partially indeterministic. His argument seems to be the following: Experience tells us that we are not the conscious source of our thoughts and choices (they merely appear to us), neuroscience tells us that those thoughts and choices arise from unconscious neural processes, and philosophy/physics tells us that these (unconscious) antecedent causes determine our present thoughts and choices. In the other direction: philosophy/physics tells us that we are determined by antecedent causes, neuroscience tells us that these antecedent causes are unconscious neural processes, and experience shows us that thoughts and choices arise out of nothingness into consciousness.

This structure of mutual support completely glosses over the multitude of assumptions that are made in the each interpretation of concept, evidence and experience. Just as Harris' claim about subjective experience is otherwise unexplained and unfounded, his interpretation of evidence from neuroscience is only one of several possible, and, as I shall argue in Part 2 of this thesis, barring his (foregone) conclusions about (meta-) physics and experience, it is not even a plausible one. Finally, although he to some extent makes explicit his stance on the question of free will and determinism in relation to other possible positions (Harris 2012, pp. 27ff), his stance is still no more than that; a stance, and he fails to tackle the substantial problems associated with it – problems to be examined here in Part 1.

The three prongs of Harris' argument only work (in unison) if his handle on the debate is granted. While I will have little to say about his personal experience, I will argue that his trident falls apart when the theoretical and empirical claims are investigated on their independent merits.

2.1.2 Making Distinctions

Whether neuroscience can decide the question of free will depends, therefore, on what you take free will to mean, what aspect of it you are discussing, and what position you hold on those various aspects. In order to evaluate claims made on the basis of neuroscientific results, it is especially important to distinguish between the theoretical question of whether free will (of some kind) is at

all possible, e.g. if it can accord with acceptable (meta-) physical theories; and the empirical question of whether humans actually have something like free will, e.g. through the role played by consciousness with regards to behaviour. In this part of the thesis, I will therefore initially be concerned with separating these two claims before I move on to discuss the theoretical question of whether free will can accord with plausible metaphysics and/or accepted theories of physics.

Among the neuroscientific research to which Harris refers in his quest to settle the question of whether humans have anything like free will is a series of seminal results published by a research group lead by physiologist Benjamin Libet. Libet's experiment will serve here as an introduction to the debate concerning free will and determinism (section 2.2), and along the way I will show why any answer to the fundamental philosophical question in this debate lies outside the limits of what neuroscience can provide (Chapter 3, especially section 3.4).

It will, however, also become clear that the kind of position one adopts on this philosophical question (potentially) affects the legitimacy of any subsequent empirical discussion on the subject of neuroscience and morality – the most obvious case being the kind of three-part denial performed by Harris (Harris). Because of this, I will briefly review the three standard responses to what might be termed the deterministic challenge to free will (Chapter 4), reviewing also one radical reinterpretation that places the acting agent at the centre of the deterministic universe in a bid to make up for the faults of the others (subsection 4.4.1), before finishing with a naturalism-inspired take on the issue which aims to diffuse the essentially sceptical worry that our beliefs and practices concerning free will are in need of external justification (Chapter 5).

But first, Benjamin Libet's attempt to time the will.

2.2 Libet's Timing of the Will

In the early 1980s, Benjamin Libet and colleagues published a series of articles on the relationship between electrical activity in the brain and the voluntary initiation of movement (see e.g. Libet and Gleason 1982, Libet et al. 1983). The surprising results of the experiments, which would garner massive attention from both scientific and philosophical communities, stemmed from the timing of the parts involved: a characteristic electrical signal leading up to a voluntarily initiated movement was detectable about 350 ms (0.35 s) *before* the time at which the subject reported choosing to move. Seeming to show that the choice to move comes to the scene only after preparation for movement has already been initiated, Libet's articles contributed to the re-ignition of the debate concerning free will and determinism, in effect opening this long-standing philosophical problem to new empirical evidence from the emerging field of neuroscience. Often claimed either to (finally)

have shown free will to be a manifest illusion, or also denied any bearing on the question, these articles remain to this day a reference for most any discussion about the empirical issues surrounding free will.

2.2.1 The Experiments

Libet's results came from a relatively simple experiment based on a well-established electrophysiological signal measurable in the motor cortex of subjects about to perform a simple action like lifting a finger. Still often referred to as the *Bereitschaftspotential* (BP) after the name given to it by the German pioneers in the field (Deecke 1965), this *readiness potential* (RP) is detected with the help of electroencephalography (EEG), a method of measuring brain activity with an array of electrodes placed on the scalp of the subject, picking up on gross changes in the electrical activity of firing neurons in the cerebral tissue underneath. Showing a slowly upward-sloping signal (increase in negative potential) which drops off rapidly at the moment of muscle contraction (as measured by electromyography of the relevant muscles, EMG), the RP is thought to reflect the unique involvement of the supplementary motor cortex (SMA) in the initiation of voluntary movements, and is relatively week in (voluntary) actions performed habitually; totally absent in those performed compulsively (e.g. by Tourette's sufferers, Libet 1998).

In the original experiment, subjects were seated in front of an “oscilloscope clock” around which the light dot of an oscilloscope would move about 24 times faster than a normal clock. Every full revolution would take 2.56 s instead of 60 s, with lines at each 1 or 2.5 “seconds” indicating the passage of 42,7 or 107 ms of actual time, respectively (Libet et al. 1983, Libet 1998). Subjects were told to flex or flick their right hand (fingers or wrists) at will, and to note where on the clock face the dot was at that time. Averaging the measurements over 40 trials per subject, Libet now had two time points in relation to which the EEG could be analysed: the reported time of initiation of movement, what Libet later calls “conscious will” or W (Libet 1998), and the physical initiation of the movement as measured by the electrical impulse leading to contraction of the right index finger. The finding was that, on average, the reported time of initiation of action was about 200 ms (0.2 s) before actual movement. But the onset of RP could be detected already 550 ms before contraction of the muscle, meaning that onset of RP on average came about 350 ms (0.35 s) before the time reported by the subjects to be the time at which they became aware of the “will” to move. As a control experiment, subjects were given an electrical stimulation of the skin without advance warning, but with the prior instruction to note the time at which they became aware of such a stimulus. On average, subjects would report the time of becoming aware of the stimulus as 50 ms before the actual delivery of the stimulus (no RP is detectable in these cases), indicating a consistent

error in some part of the report task effectively resulting in a back-dating of stimuli. Supposing that this is also valid for reports in the main experiment, the corrected time of awareness of "will" is 150 ms before actual movement, and there is a difference of 400 ms between onset of RP and awareness of the "will" to move.

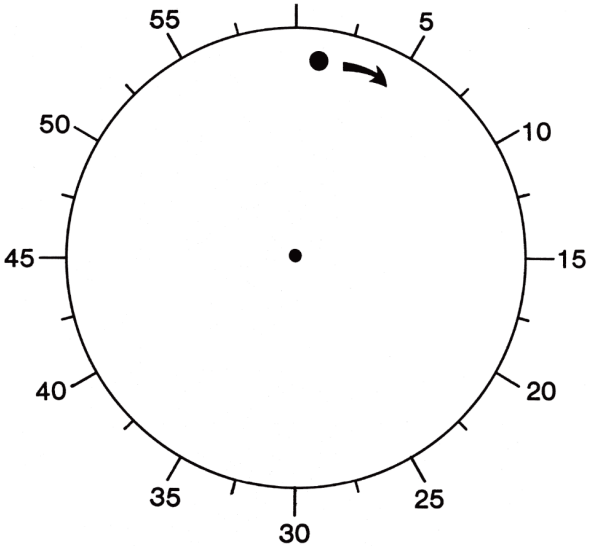

Figure 2: Schematic representation of the "oscilloscope clock" used in the Libet et al. experiments. Source: (Libet 1998)

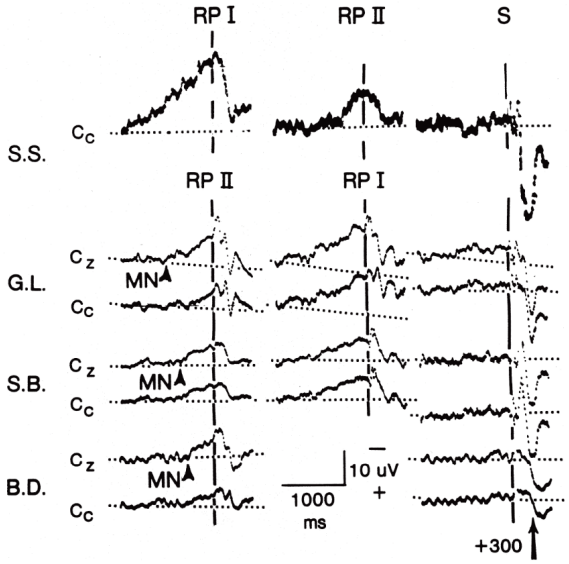

Figure 1: The averaged RPs and skin stimulation responses of four of Libet's original experimental subjects. Source: (Libet 1998).

2.2.2 Reactions

The papers published by Libet and his colleagues "engendered an avalanche of scientific and philosophical commentary" (Sinnott-Armstrong 2011, p. 8). In time, the responses have mainly fallen into one of three positions regarding the relevance of Libet's findings to the debate concerning free will and determinism: The first kind of response takes Libet to have provided empirical proof of what many philosophers and scientists until then had argued on a theoretical basis, namely that free will is an illusion. In the contemporary debate, Daniel Wegner represents one of the most advanced version of this position (Wegner 2002), and Sam Harris refers to both Libet and Wegner (Harris 2012). The second kind of response argues that Libet's experiments suffer from conceptual and/or experimental shortcomings serious enough to deny the results any relevance to the free will-debate. Adina Roskies provides a clear and concise summary of some of the critiques levelled against Libet *et al.* in her contribution *Conscious* to *Will and Responsibility* (Roskies 2011). The third kind occupies a middle position, admitting the experiments some relevance to the question of free will, while denying that they provide reason to disbelieve (in) the phenomenon altogether. Walter Sinnott-Armstrong (2011) and Neil Levy (2007) represent two different versions of this position.

While the third response – due to its recognition both of the weaknesses of the experiments, and the insights available in spite of this – is likely the most enduring, the second kind is important because Libet's experiments do indeed have serious shortcomings, both in terms of experimental protocol, and in terms of the conceptual foundations on which these rest. But to understand the dynamics that shape the debate to which all three positions are contributions, we must look into the foundations of the kind of "challenge to common sense" represented by the first response, i.e. the claim that these results offer proof against free will. In order to do this, we will have to deal with the variegated ideas and arguments around the ominous word *determinism*.

2.2.3 Determinism and the Illusion Argument

On the face of it, the claim that free will in some sense is an illusion is a very strange one. For is it not obvious that we are free to decide for ourselves what we do? Am I not now writing what I want this line to display to you, the reader, and are you not right now reading of your own free will? Granted, I am in some sense obliged to write something (namely in order to graduate), and you are probably obliged (e.g. by the need to grade this thesis) to give it at least a cursory glance, but are you not also free to choose exactly when to start reading and when to stop? I am quite sure that I could have chosen to do something else just now, or that I could simply be sitting somewhere else, typing something along these same lines, but slightly different; neither of which is necessarily incompatible

with the obligation to finish this thesis. In other words, the sum of all our obligations and preferences appear to be compatible with several different concrete approaches to fulfilling them. Furthermore, while it is trivially true that we only do a minute subset of all the things it seems we *could* do (especially if we consider all the things we could do if only we did not care about consequences or long term plans), it appears equally trivially true that there is a real sense in which we *could* actually do many of these other things, if we were so disposed.

With such evidence amply available from everyday experience, it is difficult to see why anyone would find the idea compelling that the arguably commonplace experience of being free is an illusion. To understand why this, nonetheless, is a respectable philosophical position, it is necessary to introduce what is often considered the fundamental problem for any coherent conception of free will: *determinism*. Briefly, the idea of determinism in its most general form is the claim that "every event is necessitated by previous events and conditions together with the laws of nature."(Hoefer 2010) Now, let us assume for simplicity's sake that having a free will implies, in principle if not in practice, being able to choose freely between alternatives. The problem now is that if determinism is true, free will in this sense looks to be impossible: if every event, including your choice, is necessitated by previous events and conditions together with natural laws, it seems that you have no real choice after all. Whether you think of it as there being no real alternatives, or if you think that the choosing itself is "false" or illusory, determinism appears to render impossible any such notion of freedom. Still, we *feel* that we choose "freely" in some sense, that we have the ability to do as we see fit, and not simply play out a series of events necessitated by previous events and the laws of nature. This is where talk of illusions comes into play through what may be called the "illusion-argument":

> IA: Our *experience* of (free) will is nothing more than that, an experience, and it does not tell us the truth about what is going on. Our choices and actions are determined, and the experience of free will is therefore illusory. (Wegner 2002)

The Libet *et al* experiments have been taken by some proponents of this position to be the first empirical evidence in support of the illusion-argument (Banks and Pockett 2007). RP was, more or less explicitly also by the original authors, interpreted as the physical precursor of voluntary action, *the real deal* in terms of how behaviour is initiated or caused (Roskies 2011, p. 15), and the report by subjects of an awareness of an "urge" or decision to move, while initially associated with the "intention" to act (Libet et al. 1983), has later been identified with "conscious will" (Libet 1998). Under this interpretation, the evidence appears clear: conscious will (or intention) comes to the scene too late to be the origin of movement. Early critics of this interpretation attacked the various

methodological and conceptual shortcomings of the experiments, and these do indeed have serious consequences for the kinds of conclusions licensed by the results. However, other research groups have in the years since performed variations of the Libet *et al.* experiments addressing these issues and testing new hypotheses related to the role or importance of the conscious experience of initiating action. Among these are experiments successfully manipulating the reported time of "conscious will" using magnetic manipulation of the pre-supplementary motor area (preSMA, one of the areas of the cortex involved in voluntary behaviour) applied *after* movement itself was begun (Lau, Rogers, and Passingham 2007). Another experiment achieved similar results using an auditory beep instead of magnetic stimulation (Banks and Isham 2009), strengthening the claim that our experience of wilfully initiating action is not related to movement in the straightforward way of an immediately preceding cause or initiator (Sinnott-Armstrong 2011, p. 13 ff.). Finally, John-Dylan Haynes *et al.* have published a remarkable Libet-style experiment in which they predict not only that action will be performed, but also which of two alternatives will be chosen, with information about which of two options will be chosen available from data recorded up to 10 seconds before the subject acting is aware of making a choice (Soon et al. 2008b, Haynes 2011a).

With these experiments giving such intuitively strong support to the illusion-argument, the assumption that they also bear on the question of determinism itself is perhaps also tempting: the experiments may seem to provide clear evidence for the idea that every event is the result of antecedent conditions sufficient for bringing it about, with the "event" of a hand movement being the result of the preceding RP. While intuitively appealing, the thought is false: Libet's (and Libet-style) experiments have no bearing on determinism *as such* (Sinnott-Armstrong 2011, p. 2). In order to understand why this is the case, we first need to understand more fully what determinism is and what it is not.

3 Free Will and Determinism

> "To have free will is to have what it takes to act freely. When an agent acts freely—when she exercises her free will—what she does is up to her. A plurality of alternatives is open to her, and she determines which she pursues." Randolph Clarke (2009, p. 1)

> "The term 'free will' is a philosophical term of art. […] The first thing to realize about the use of the words 'free will' by philosophers belonging to the classical tradition is that, now at least, these words are a mere label for a certain feature, or alleged feature, of human beings and other rational agents, a label whose sense is not determined by the meanings of the individual words 'free' and 'will'. In particular, the ascription of "free will" to an agent by a current representative of the classical tradition does not imply that the agent has a "faculty" called 'the will'. […] When a current representative of the classical tradition says of, e.g., Mrs. Thatcher, that she "has free will," he means that she is at least sometimes in the following situation: She is contemplating incompatible courses of action A and B (lecturing the Queen and holding her tongue, say), and she can pursue the course of action A and can also pursue the course of action B." Peter van Inwagen (1989, p. 400)

3.1 Historic Background, General Classification

The debate concerning free will and determinism can trace its two main roots back as far as the pre-Socratic Atomists of 5th century BCE Greece and to the shift from polytheistic to monotheistic religions in ancient Mesopotamia and Greece (Eshleman 2009). It is probably also related to the even earlier reflections around *fatalism*, or the thought that (some part) of ones future is predetermined (e.g. by the gods) in such a way as to make all one's own deliberations and actions irrelevant to whether that fated future is realised (Eshleman 2009).

Arguments that have their roots in the early monotheistic religions are grouped today under the heading "theological determinism". One of the better-known arguments for theological determinism is the problem of "divine omniscience" or a god's foreknowledge. The problem, in brief, is this:

> Suppose that being omniscient entails being infallible, believing p if and only if p is true
>
> Suppose that an omniscient god existed in 1900
>
> Suppose, further, that Jones mowed his lawn on 1/1/2000
>
> Then the omniscient god believed in 1900 that Jones would mow his lawn on 1/1/2000
>
> It follows from this that Jones could not do other than mow his lawn on 1/1/2000

This follows since any other action would either 1) make it true that the god held a false belief in 1900, 2) make it true that the god did not hold this belief in 1900, or 3) make it true that the god did not exists in 1900 – all of which have been ruled out by the assumptions. Thus, if there is an omniscient god, it appears that we are determined to do as we do (Rice 2013, Pike 1965).

"Scientific determinism" is the modern counterpart to the determinism of the Atomists, and shares with theological determinism the fundamental idea that, given a certain starting position, things can only turn out one way. Where theological determinism attributes this to the nature of a god, scientific determinism sees this as a logical necessity arising out of the combination of prior states of the universe and natural laws governing the transformation of these (Eshleman 2009). A classic formulation of such *nomological* determinism is due to Peter van Inwagen:

> If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (Van Inwagen 1983, p. 56, quoted in Vihvelin 2011)

In the following, I will exclusively deal with the scientific species of determinism, and all subsequent references to 'determinism' are therefore references to this.[4]

3.2 Subspecies of Scientific Determinism and Single Cause Theories

In addition to these fundamental distinctions, some classifications include a number of "special" scientific determinisms like biological (Lewontin 1982) and psychological determinism which have gained currency with the increased sophistication in theories e.g. about the roots of human behaviour in evolutionary adaptation and the psychological mechanisms which often work beneath the level of awareness to produce behaviour of some predictability (and in a possibly more sinister vein, manipulability). To these one might also add a neuroscientific determinism, in which the issue of determination could be seen to arise from the project of explaining mental phenomena with purely physical descriptions of deterministic processes in the brain. In another sense, these "special" determinisms are unlike the fundamental scientific determinism in that they offer empirical arguments against the thesis or idea that humans have (something like) free will in any robust sense that they are likely to practice in relevant situations, rather than argue that the concept itself is incoherent or impossible in principle or given fundamental facts about the laws of nature. Thus, a social psychologist may do a series of experiments on the *prima facie* irrelevant features of situations

[4] Note also that most if not everything that is true of determinism in this context also extends to plausible forms of indeterminism, e.g. the kind espoused by the Copenhagen-interpretation of quantum mechanics, according to which the world is fundamentally probabilistic.

that actually turn out to "determine" choices of moral import, and conclude that situational features far outweigh anything like character in deciding how we act.[5] These can therefore also be seen as "major factor" determinisms, i.e. claims about which factors *most* influence thought or behaviour. Sometimes, such theories slip into the extreme end where other factors are considered irrelevant, and a single cause is picked out as that which determines behaviour, etc. The two extremes of the "nature/nurture"-debate illustrate this point, where dogmatic positions – we are either fully determined by genes, or fully determined by society – obscured the now-recognised importance of both biological and environmental/cultural factors in the shaping of our thoughts and behaviour. One of the reasons why the nature/nurture debate became so heated (a fact to which Harris' quote at the start of this chapter in part refers) was because it appeared that the new science of *socio-biology* – later to evolve into evolutionary psychology – was trying to establish the evolved biological determinants of human behaviour. In its least nuanced form, human behaviour was reduced to the "self-interest" of our genes: whatever behaviour ensured the spread of corresponding genes to the next generations would flourish (Dawkins 1976). Many were unhappy with this assertion, as there seemed to be little freedom as a puppet of our genes. At the opposing extreme was the idea that human beings were born as "blank slates" onto which any culture or behaviour could be inscribed. At least part of the allure of the "standard social science model" (Barkow, Cosmides, and Tooby 1992),must have been that it seemed to ensure a healthy freedom to decide for ourselves what to do and whom to become: A blank slate is waiting to be filled in with what we want it to display. The problem with this is of course that the idea was part of social determinism, and as such offered no more freedom to the individual than did its biological cousin: instead of being determined by genes, our nature, it was society, our nurture, that made us into what we are.[6]

Both of these extreme positions are recognised today as examples of the "single cause" fallacy (de Melo-Martín 2005), mistakenly singling out a single cause as being responsible for something which in fact is the result of the often highly complex interaction of a plethora of factors. But note also that this does not affect determinism as such in the slightest, as we are not discussing *whether* behaviour is determined, only *what* the determinants are. Interestingly, such "single cause fallacies" precipitate the fear of determinism in a way plural-cause accounts appear not to. It may be that picking out a single cause for our behaviour makes vivid the claim that we are wholly determined,

[5] Thus, certain experiments – showing e.g. that finding a coin in the coin slot of a public telephone was by far the best predictor of subsequently helping someone who drops all their documents on the floor – have motivated a position in psychology known as "situationism". Situationists argue that our actions are far more influenced by situational features than determined by anything like stable character traits. See e.g. (Harman 1999, Doris 1998).
[6] As Stephen Pinker has put it: "a blank slate is a dictator's dream", seeing as how it would allow a dictator to inscribe whatever system of repression she could wish for onto the "slates" of her subjects.(Pinker 1994, p.427).

even if it is 1) a fallacy and 2) no more deterministic than determinism in general.[7] Perhaps we think that if there are many and murky determinants, one of them (and possibly one of the most important) must be our own (free) will. While this might seem like wishful thinking to the burgeoning sceptic, there is another sense in which we are absolutely right in thinking that our own "will" is a major factor in determining our behaviour, namely as one of the important *causes* of behaviour.

3.3 Causes of Behaviour

Somewhat surprisingly then, determinism is quite compatible with the commonplace observation that our choices are what might be called *causally effective*, i.e. that we tend to do the things we choose to do. This is "surprising" because a lot of talk about determinism portrays it as a doctrine in which human choice and action is to be seen as controlled by forces outside of our control. Determinism is usually taken to sideline the faculties by which we normally think we navigate our lives, like rational deliberation, evaluation, decision, etc. Instead, thoughts and behaviour are seen as caused by e.g. our brain, our genes, the environment, our upbringing, society, or simply the state of the universe at some point in time long before our births – in short, causes external to the agent acting.

However, determinism *per se* provides no reason for limiting determining factors to those that are external to the agent acting, nor any grounds for external factors to be emphasised in explaining behaviour. To remind ourselves, the fundamental claim of determinism with which we are here concerned is simply that "every event is necessitated by previous events and conditions together with the laws of nature."(Hoefer 2010) There is no clause specifying what role different factors or causes play, simply because determinism as such does not concern itself with the individuation of factors or causes. Determinism is actually more counterintuitive than you might think from reading popular discussions of the subject. If you picture the world according to determinism as a vast box also stretching out in time (which would require four dimensions if it were to be pictured all at once, so just let it play out freely in any "direction of time" in your imagination), no single point within that box (including any point at any "point" in time, i.e. any four-dimensional coordinate) is any more important than the others.[8] No cause or factor that would be singled out as important by a human

[7] The fact that ideas of fate and fatalism are related to the rise of monotheistic religion is an interesting historical parallel to the "single cause" fallacy. A single God appears to present problems for our thinking that a plethora of gods and demi-gods do not, and identifying a single cause of behaviour intuitively threatens freedom in a way a multitude of factors appear not to.

[8] While this representation of time and space as a four-dimensional block (known as "block universe") is controversial, so are its alternatives (i.a presentism and growing block universe). For more on this, and especially on the relationship between special relativity and space-time, see (Petkov 2005).

observer occupies any especially significant position *from the point of view of determinism*. This has to do with the way we understand the world to be according to our best theories in physics. Assuming that those are approximately true,[9] the world is a vast network of "stuff" which can be considered reciprocally interacting in either wholly or partially deterministic ways. It is not the case that there is one kind of "thing" (a cause, say) which exclusively determines another (an effect), and therefore is the appropriate place to locate the kind of "responsibility" inherent in the claim that your behaviour is determined by your genes or the kind of socioeconomic environment in which you grew up. Neither of these are the end of the line as far as determinism goes. Your genes did not arise out of thin air to serve as the uncaused cause of your behaviour, and the socioeconomic environment of your behaviour is just as much an effect of other people's actions as it is the cause of any of yours.[10] And you in turn shape them both.[11]

Determinism can get even more counterintuitive when we consider the time-symmetry of physical laws. I will get back to this in section 4.4. For now, the upshot here is the important clarification that most talk about determinism and free will brings into it something that is not inherent in determinism itself. One of those things is the intuition that external causes become more or even exclusively important in explaining behaviour if determinism is true.

Even if this intuition cannot find direct support in determinism, it is, at least to some extent, understandable as a reaction to this, in so far as determinism also does not *privilege* those faculties that traditionally have served as the endpoints of explanations of human freedom, either. To the extent that determinism undermines these faculties as they are traditionally understood, the drive to look elsewhere for an explanation of both action and the (apparently false) experience of freedom, is perhaps reasonable.

This is also where empirical human sciences such as neuroscience can enter the picture: with determinism having undermined traditional explanations of free agency, science can provide

---

[9] This is the least controversial version of *scientific realism*, a variant of realism about the existing world according to which our best current physical theories about the fundamental nature of that world – and all the macroscopic features of the human world with which these are compatible – are at least *approximately* true, i.e. not completely, hopelessly wrong as descriptions of the world as it actually is. For a comprehensive treatment of varieties of scientific realism and its alternatives, see (Dudau 2002).

[10] I am shifting between quite different levels of description here, going from the kind of micro physics described by physical laws to the macro world of human organisms and social interaction, but the point is (I believe) transferrable: take away our heuristic interpretations and valuations, and what you are left with is the pure claim of determinism that everything is necessitated by law-governed changes between states of affairs.

[11] That you shape your socioeconomic environment is trivially true in the sense that this environment is the result of complex interactions between individuals, institutions, etc., some of which are interactions directly involving you. In case of genes, it is (mostly) the *expression* of your genes that is affected by interactions with the environment in which you as an agent take part, but since genes themselves can claim no primacy (neither explanatory nor physically) to the pattern of gene expression (epigenetics), I allow myself the simplistic formulation used above.

evidence for the move to relocate the nexus of control outside the now supposedly discredited notion of a freely choosing and cohesive "agent self".

Very likely, there is a confluence of interest coming from opposite directions on this issue: philosophers (and others) convinced by determinism of the falsity of traditional conceptions of free will, find confirmation of their views in contemporary science casting doubt on the importance of a conscious agent, or even a coherent "self", in action initiation and control. And scientists investigating the "real" mechanisms behind such fuzzy concepts as action control and "conscious will" find support for wholly "lower level" deterministic physical interpretations of these phenomena in the (separate) theoretical claim that everything necessarily is deterministic in this way. This seems to be the case, e.g. for Sam Harris's approach to the debate, as discussed above.

However, these matters are in principle independent, and they must be criticised on their separate merits. The slide is easily made and often difficult to notice, and it is therefore important to note that determinism does not licence or even support conclusions about the locus of action control (or something like it). Indeed, determinism is compatible with several different accounts of action control, some quite surprising in the power and freedom they can afford an acting agent.[12]

That being said, the position you adopt on determinism does have consequences for what can be said about human freedom and morality, as we shall see presently. I therefore think it necessary to give a very brief overview both of the common positions taken on this question, and to present a slightly more radical alternative which helps undermine some of the more intractable problems which face anyone arguing for a scientifically supported conception of freedom.[13] Finally, I will explore the option of regarding determinism's "challenge to common sense" as a sceptical challenge similar to the classical, epistemological sceptical challenges, in a bid to diffuse the worry that free will is somehow ([meta-] physically) impossible.

3.4 One Debate at a Time, Please.

Free will is hard to define. So far, I have mainly deferred to the reader's own understanding of the concept, adding here and there bits of specifications which have been given prominence in the debate surrounding the relation between free will (or freedom) and determinism (and

[12] I will present one of these accounts in section 4.4.1 where I talk of Hoefer's "Freedom from the Inside Out" (Hoefer 2010). What exactly is meant by "freedom" here is of no great consequence, as my argument in this part of the thesis does not hinge on any positive account for freedom/free will within current interpretations of determinism, but rather on a sort of "discrediting of the sceptical challenge" that determinism can be said to represent. See (McKenna and Russell 2008) and discussion here.

[13] Due to place constraints, I will have to ride somewhat roughshod over the very great variety of positions and arguments in the classical debate, and will certainly do them injustice in this. I hope this can be excused on the basis that this thesis is not concerned with adjudicating between classical responses to determinism.

indeterminism), and clarifying some implications of determinism for this. For the purposes of this section, van Inwagen's remark (quoted above) on the "classical tradition" is a suitable definition of what is meant by "free will" in the kind of fundamental conceptual debate with which we are here concerned. That free will is not considered a particular faculty and is left unspecified in cognitive, physiological or other scientific terms is also noteworthy in relation to the distinction made above between the classical, philosophical approach to the "Determinist Question" (Kane 2005, pp. 7ff) and the way the question of free will is dealt with in scientific studies on human behaviour and decision making. In the latter case, one often talks of action control (or initiation, depending on the conceptual framework) and whether or how this is exercised when someone chooses or acts. That these are separate issues is clear enough: the fundamental, philosophical/theoretical question is whether anything like free will is *possible*, while the scientific or empirical question is whether humans *actually have or exercise* anything like free will. Thus, Libet's experiments investigated whether the conscious intention to move a finger was the actual cause of the finger moving, and found, for these cases, the conscious intention to move (W) to be preceded by brain activity (RP). Whatever conclusions this might be taken to licence about the relationship between conscious intentions (or will) and action, it does not speak to the question of whether every event – including the events of conscious intentions and overt actions – is determined by preceding states of affairs together with natural laws. Dealing with determinism here, I am therefore engaged in clarifying whether something like free will is *at all* possible, and whether it is possible in principle for something like a human agent. All questions of whether any actual humans *are* candidates for having free will and whether any humans *have* or *exercise* free will are deferred to Part 2.

4 Responding to Determinism

> "There is a disputation that will continue till mankind is raised from the dead, between the necessitarians and the partisans of free will." Jalalu'ddin Rumi, twelfth-century Persian poet and mystic (quoted in Kane 2005, p. 1)

There are at least two general ways of responding to determinism in the context of the kind of philosophical discussion of free will that we are undertaking here. By far the most common type of response accepts that determinism and/or (plausible) indeterminism poses a challenge for traditional conceptions of free will, and argues for the consequences this has or does not have for the way we regard each other and ourselves, especially in terms of moral and legal responsibility. The three traditional responses to determinism – *hard determinism/impossibilism*, *libertarianism* and *compatibilism* – all fall in under the first type of response, in that they all accept that determinism and/or plausible indeterminism poses a *prima facie* challenge that must either be accepted (hard determinism/impossibilist), repudiated (libertarianism), or incorporated into a revised notion of free will (compatibilism).

The second, far less common type of response is to deny that determinism poses an appropriate challenge to our practice of considering each other free and holding one another responsible for the things we do. This is not only a minority position, but represents a completely different way of relating to the debate concerning free will and determinism, which is why I will refer to it as a "fourth option". Because of this relation to the classical debate, I will return to the second kind of response only in Chapter 5 , dealing first with the consequences the acceptance of determinism can have, then the different positions that have arisen in the discussion of these consequences.

In order to understand the felt need for these kinds of responses, we should first of all clarify *how* determinism can pose a threat to free will as defined above; secondly why this threat is considered important enough to justify such a voluminous debate.

4.1 The Real Threat from Determinism

Suppose that determinism is true. Things can only turn out one way. There are practical and principal reasons for thinking that we will never be able to tell which way things are going to turn out (apart from in the very vague, error-prone and short-sighted guesswork we currently employ).[14]

[14] It's important to keep the question of predictability apart from that of determinism, especially as it arguably is the contemplation of the possibility of predicting our future actions with 100% certainty that provides much of the intuitive drive to our fear that determinism precludes freedom (Holton 2013). While I consider it an open question whether determinism entails predictability *in principle*, it is at least overwhelmingly likely that predictability is impossible for any "thing" of which it makes sense to say that it can "know" the future. There are two principal and one practical reason for this: The practical reason is that complex systems can develop in highly divergent ways based on tiny differences in

It is plainly obvious that how things turn out depends on your actions, and if anything, the truth of determinism supports this observation, as noted above. Therefore, you had better act so as to bring about the world you would most like to come into being. Now, what is added by saying that you are determined to do this? Does it entail that *you* are not doing your best to make the world a better place? Not at all, for even if one allows that your actions are determined, this does not entail that someone *else* is doing these things for you, nor does it entail that no one is doing anything.[15] Does it entail that your actions are useless in bringing about this brighter future? No, your actions are absolutely essential to bringing this future about: we may very well say that the future would be different were it not for your actions. So where's the rub? The rub, it seems, lies in this: determinism takes away our ability to do otherwise. That is, one arguably central part of our notion of freedom is the thought that we are free to choose what to do now; we have real alternatives, and at the end of the day, it is up to us to choose. Determinism seems to undermine this by *necessitating* that you choose and act the way you do. However much you identify with or find abhorrent your behaviour in a particular situation, being the physical being you were in that situation, you could not have done otherwise. The important thing to keep in mind here is that the position which espouses this view (hard determinism) not only intends that you, given a particular situation, due to the kind of person you are, your values etc., are overwhelmingly likely to choose a certain *kind* of course of action, e.g. that you are more likely than not to give a moderate sum of money to a homeless person in a particular instance. Rather, there is only *one* unique course of action "open" to you. You cannot even choose *not* to follow that course of action, and hence, you do not have an ability to do otherwise.

---

starting conditions. The two principal reasons tie in with the practical one: Heisenberg's uncertainty principle states that for any single particle, only one of a pair of properties can be defined with precision at a time (a fundamental feature of the wave nature of matter according to quantum physics), meaning that a particle cannot both have a precise position and a precise momentum at a given time. The second principal reason to keep predictability separate from determinism is that any observation of a system will change the system (what is known in physics as the observer effect), and you must therefore include the measurement into the system it measures. This creates a host of difficulties for accurate prediction, since e.g. for quantum phenomena, the state of a system can depend on whether it is observed or not (Schrödinger's cat is an example of this, see e.g. Gerrits et al. 2010).

[15] This distinction should be kept in mind when reading thought experiments in which some sinister puppet master is posited to ease the notion of determination into our intuitive view of responsibility (as is done e.g. by Greene and Cohen 2010). When at the end of the thought experiment the puppet master is exchanged for impersonal determinism, our intuitive response to the thought that someone else, another agent, is controlling us is supposed to transfer onto the idea that our actions are determined (and easily does). This, however, is an illicit bait-and-switch, since in the first instance our intuition that we are unfree when controlled (arguably) arises out of the specification that we are being controlled by *someone else*, someone with agency and responsibility, allowing the transfer of responsibility onto this other agent. Determinism is unlike this because no one is "being controlled" at all. Determinism also does not entail that we are all coerced into doing what we do, nor that we are thus compelled. The only thing determinism entails is that we are determined, and even this, as we shall see later on, can be given a different perspective.

4.1.1 The Relation to Responsibility

To hold and be held responsible for the actions we perform is a pervasive feature of human interaction, most commonly and importantly applicable to the actions we recognise as performed of a person's "own free will". Given this idiomatic gloss, it is easy to see how a "free will"-denying determinism might challenge our common practice in this area. Indeed, most discussions of determinism and free will present the challenge from determinism not simply as a challenge to free will, but immediately also note that this would seem to undermine any notion of moral responsibility, and therefore, in turn, the basis for our practices of praise and blame, reward and punishment. This is one of the most important reasons why determinism is so widely debated in connection with free will. If our practices of moral (and legal) judgement are inconsistent with our best knowledge, this seems to present a very strong case for the revision of common practice. If people are not really free in the way we think they are, it seems highly problematic to keep on treating them as if they were, and not only from the point of view of intellectual consistency. People are blamed and punished for their actions every day, and even the most advanced legal systems can arguably be said to perpetuate some trace of the notion that the people thus punished *deserve* their punishment. On the other end of the scale, we regularly applaud and otherwise revere those who have made something of themselves or done something heroic or great. We certainly tend to think that they deserve praise for their actions, and while there are few pure meritocracies, the notion that rewards should be *earned* has broad acceptance across the globe.

These are, importantly, issues of perceived fairness, social aspiration and a multitude of other phenomena that can be analysed purely descriptively in humanistic and natural sciences. But we may also consider them normatively, in terms of right and wrong, independently of the motives that actually move people in practice. It seems intuitively obvious that the possibility of free will is a requirement for the rational appropriateness of holding people responsible, for good and bad. After all, if what you did was determined to happen by the state of the universe at some time in the distant past together with the impersonal laws of physics, how could *you* be considered responsible for doing it? If someone can be said to have acted knowingly, or knowingly failed to act, we typically only excuse their (in-) action if any one or a combination of four things holds true: they were either physically restrained (e.g. tied up), coerced (e.g. at gunpoint), lacked the ability (e.g. because paralysed), or lacked the opportunity (e.g. because on the other side of a heavily trafficked street). If it is true that what you do is determined to happen by the state of the universe at some time long outside of your reach, then, if you are not literally coerced or compelled, it *seems* that something very like it is taking place. You had no choice; you *could not have done otherwise*. As we saw above,

determinism can be seen to threaten just this intuitively essential feature of responsibility, but it is misleading to compare our actions under determinism with those that we can recognise as compelled or coerced; these are distinct notions: To say that your actions are determined does not entail that someone else is coercing you to perform them, nor that there is some psychophysical force beyond your control which compels you. This is how we usually picture "unfree" actions, but the threat from determinism is of a different, in one sense more fundamental kind (or it appears to be, but as we shall see in section 4.4 and chapter 5, just what this means is far from clear).

In any case, to the extent that determinism eliminates alternate possibilities, there is a powerful intuition that this also makes traditional conceptions of moral responsibility untenable. Variations of this intuition almost always accompany the debate around free will and determinism, and while there are many ways of analysing just what it entails, the common feature of all these is the thought that the truth of determinism, in one way or another, is incompatible with free will. As a result, the position that holds this to be the case is known as *incompatibilism*.

4.2 Incompatibilism

Incompatibilism, then, is the claim that *if* determinism is true, there is no free will, and if there is no free will, most incompatibilists argue, we also have no grounds for holding people (morally and legally) responsible for their actions. We generally only hold people responsible for things they themselves chose to do and could have avoided doing if they had chosen otherwise. If we cannot hold people responsible for choosing their actions, one important reason to punish also disappears. After all, we usually do not punish or punish only mildly those who were not in a position to "do otherwise", e.g. because of coercion by an external force, or through a compelling internal force outside of the agent's control. In all such cases, it is arguably the fact that persons thus affected through no fault of their own had no other (or no other real, available) choice than to perform the act for which we are then liable to excuse them. If determinism is true, then i) there is no free will, therefore ii) people have no "ability to do otherwise", which in turn means that iii) we have no grounds for holding them responsible for their actions in this sense, and lastly, iv) at least one justification for punishment – "just desert" or *retribution* –disappears; or so runs a familiar incompatibilist argument.

Among the incompatibilists, one traditionally finds two contrary positions: *hard determinists* and *libertarians*. Hard determinists claim that determinism is true, and therefore that free will does not

exist. Libertarians claim that free will exists, and that determinism therefore must (at least in part) be false.[16]

4.2.1 Hard Determinism

Hard determinism is, as already pointed out, controversial just because it claims that humans do not have anything like free will. Since there is no such freedom, we also cannot hold anyone responsible with the justification that they could have done otherwise. To the extent that we actually depend on such justification in holding persons responsible for their actions, we face a serious inconsistency between our practices and what this reflection tells us is true. On the basis of this, many hard determinists conclude that we need to amend our practices, typically by completely removing any reference to *retribution* in legal contexts (Greene and Cohen 2010, Harris 2012). However, few hard determinists want to release all criminals, and so they are quick to point out that lawmakers have another option, namely to punish solely with regard to the *effects* punishment is likely to have on the criminal her-/himself and society in general. Such *consequentialist* reasons are already prominent as justification for punishment in legal systems based on the Roman tradition, and it is common to speak of deterrence, denunciation, incapacitation, rehabilitation, and reparation as aims for punishment – justification for which can be formulated without any reference to whether or not the criminal could have done otherwise. Still, our everyday notions of responsibility, praise and blame *do* seem to depend, and depend to a large extent, on the idea that people really are responsible for their actions in the sense that they could have avoided – or could have failed to accomplish – those things for which we blame or praise them, and that this depends not on the state of the universe at some point in time, but somehow on those persons themselves – their baseness, courage, mettle, indecision, kindness, cruelty, wisdom, vileness, egotism, strength; or any of a number of other adjectives with which we are apt to describe the people we admire or despise. If nothing else, hard determinism seems to render all these (at best) unfounded and illusory. In complement to this other-regarding aspect of free will is the thought that we by our actions are able to change the course of the world; decide whether something turns out one way or another. While hard determinism is compatible with our actions being essential in bringing about a given future, it also denies that we have any real ability to bring about a *different* future from the one entailed by the

[16] Thirdly, there is a position known, i.a. as "hard incompatibilism" or, as already mentioned, "impossibilism", which simply adds to the hard determinist position the claim that free will (or ability to do otherwise) is incompatible both with the truth of determinism *and* the truth of indeterminism, i.e. impossible either way. Because both libertarianism and compatibilism, if successful in their arguments, can serve as a refutation both of hard determinism and of hard incompatibilism, I will not treat the latter separately here, following instead the convention I have already adopted of intermittently pointing out that what we are discussing is also valid for (plausible) versions of indeterminism.

given past. Same past, same future, as Robert Kane summarises the issue (Kane 2005, p. 16), thus rendering also this "power" illusory.

There is a tension in the hard determinist position between, on the one hand, the radicalness of the claim that there is no free will and the consequences this should have for our practices of moral and legal judgement; and on the other, the need or desire to accommodate more "everyday" notions of choice, responsibility and freedom – things it is very hard to understand how we should live without. It is not uncommon for hard determinists to submit that we cannot or should not abandon practices of moral and legal blame (and their attendant beliefs) that are unsupported by determinism, either 1) because doing so would wreak havoc on society, or 2) because people invariably will look at the world in this way, whatever you tell them. While there might be some grain of truth to such appeals to consequence (or lack thereof), they are not borne out quite so dramatically in practice: firstly, the argument for hard determinism has been known for a comparatively long time, and society still seems to be hobbling along as before; and secondly, there is scientific evidence that people influenced by the argument for hard determinism behave worse than those not under such influence (Vohs and Schooler 2008). In other words, it is neither true that people are unaffected, nor true that the effect is that of Armageddon.

One option to try to ease this tension is to see whether we actually need to posit free will and moral responsibility to make sense of or justify our practices. I have already mentioned how one can justify punishment without recourse to notions of desert, and Derek Pereboom deals with this and related issues in his book *Living Without Free Will* (Pereboom 2001). According to Pereboom, the only thing we are really giving up if we accept hard determinism is the thought that we are *ultimately*, in the sense "at the bottom of it all", responsible for who we are, what we do and what we achieve. Not only is this a small loss, he argues, it can even serve as a source of happiness and positive societal change: accepting that we do not have (ultimate) power over what will happen to us, we can both become less dependent for our happiness on the vagaries of life (a thought with roots in the Stoic tradition of ancient Greece), and become more forgiving of the actions of others. Some traditional thinking about our own and our fellow's responsibility and moral "creditworthiness", as well as our hopes for an as-yet fully undecided future, become untenable with the loss of free will, but most of what is important to us can either be transferred without problem, or also replaced with analogues that do not have unsupportable metaphysical entailments. Pereboom quotes an instructive passage from Bruce Waller's *Freedom without Responsibility* (1990) on guilt seen without (ultimate) moral responsibility:

> It is reasonable for one who denies moral responsibility to feel profound sorrow and regret for an act. If in a fit of anger I strike a fried, I shall be appalled at my behavior, and profoundly distress that I have in me the capacity for such behavior. If the act occurs under minimum provocation, and with an opportunity for some brief reflection before the assault, then I shall be even more disturbed and disappointed by my behavior: I find in myself the capacity for a vicious and despicable act, and the act emerges more from my own character than from the immediate stimuli (thus it may be more likely to recur in many different settings), and my capacity to control such vicious behavior is demonstrably inadequate. Certainly, I shall have good reason to regret my character – its capacity for vicious acts and its lack of capacity to control anger. (Waller 1990, pp. 165-6, quoted in Pereboom 2001, p. 205)

In other words, much of what matters to us can arguably be said to be independent of a true notion of ultimate responsibility or free will in the philosophical sense. However, Pereboom also warns that we cannot remain completely unfazed by determinism, as it does indeed have consequences for some of the things practiced today, most notably, and as already noted, retributivist justification for punishment, as well as some aspects of our moral judgements and emotional responses such as moral indignation.

4.2.1.1 Minimal claims and their potential consequences

There is a critical point to be made here. The Norwegian legal system is, in letter if not always in practice, free from reference to "just desert" or retributive justifications for punishment. This, however, has nothing to do with the truth of determinism, but is the result of a rejection of the thought that punishment has any inherent value (i.e. as revenge), and a shift in focus towards punishment as a means to shape behaviour to be in accordance with the laws (NOU 2002: 4). It is important to keep in mind that much of the debate concerning free will and determinism, especially when it comes to moral and legal responsibility, is shaped by a particular, Anglo-Saxon conception of desert and law, and that one may reject retribution on other grounds than the truth of determinism. This also illustrates a broader point about the need to distinguish between the minimal claims made by different positions in this debate, and the various consequences than *can* be drawn from them, but which may equally well be denied or associated with a different starting position.

4.2.1.2 A side note on an apparent paradox

There seems to be a practical inconsistency or outright paradox in the notion that the truth of hard determinism (or hard incompatibilism) should be treated as important and possibly acted upon. Say, for the sake of argument, that determinism (or indeterminism) is true in the way it is conceived of

according to the debate concerning free will and determinism. Some people see this truth; others do not (as evinced by the debate itself). However, you have no real choice about whether or not you see this, nor are you responsible or deserving of credit or blame for your position. Moreover, if it *appears* that you have a choice about whether or not to implement changes on the basis of this truth, in so far as your choice and the question of implementation of change is necessitated by antecedent states of the universe and governing laws, there exists no real choice in this matter either. Insisting, on the backdrop of these connected realisations, that we should all embrace this truth and reject moral responsibility – reforming in the process the penal system and our own, interpersonal interactions – seems downright absurd: in order for the notion of "should" to apply to the question of implementing a disbelief in free will, it must be possible either to do this *or not*; there must be a real choice about it. However, since the idea and entire motivation for modifying anything based on hard determinism is that no such choice exists, the claim appears to be self-refuting.

In other words, hard determinists may very well say that humans do not have free will, but when it comes to the question of what we believe, how we live, and how we shape society, the idea cannot carry any force in and of itself – in contrast to the force inherent in a belief in free will. This is not to say that our lives and society might not be shaped, or even shaped to the better, by assenting to the hard determinist position. It is just that there is no force in the position itself to motivate such change, and that any change thus justified is actually without justification. If any sense were to be made of such a move, it would have to be by appeal to the positive consequences of the acceptance of this (absurd) position, and as I have intimated above, any positive change supposed to follow from hard determinism can be motivated by other means without attendant absurdity, and so there seems to be no reason to shape society on the basis of hard determinism.

This appears to leave the hard determinist only with the option to say that we do not have free will, but that no consequences can or should be drawn from this realisation. I am ignorant of any solution to this paradox.

4.2.2 Libertarianism

Libertarianism, here libertarianism *about free will* and not to be confused with its political namesake, starts with the opposite assumption from hard determinism. Rather than accepting that our understanding of the world implies that freedom is impossible, libertarians takes our perception of freedom as given, and concludes that the picture presented by determinism must somehow be false (of our world). By denying determinism in part or in general, the libertarian opens up space within which we can "will", i.e. choose and act otherwise. However, by opening up this space or

gap, the libertarian is at the same time creating a problem for herself, namely how to account for a "willing" that bridges this gap in the otherwise whole cloth offered by the deterministic picture. The primary problem for Libertarianism is therefore to provide an alternative explanation of the basic nature of the world, or of that part of the world that is made up of our (free, undetermined) actions.

In other words, positing indeterminism only gets the libertarian halfway to free will, since the element of chance thus introduced is one over which we appear to have no control – indeed, such an element of randomness is a *prima facie* detriment to free will, in that whatever control attributable to the agent in the wholly deterministic picture would seem to be degraded by randomness; replaced in part by mere luck. In order to account for the kind of free will that would entail an ability for agents to do otherwise (and not simply a freedom of randomness), the libertarian must therefore give an account of how an agent or an agent's choice or decision is what brings about behaviour without being itself determined.

To account for this, libertarians typically follow so-called *extra factor strategies* (Kane 2005, pp. 38ff), where one attempts to provide an explanation of the "extra factor" that would guarantee or explain (the possibility of) free will within the opening created by (partial) indeterminism. Among the most recognised extra factor strategies are variations on *agent causation*, in which the extra factor typically is the *agent as entity*, undetermined in itself, but deterministically connected to the actions performed (Chisholm 2003, Taylor 1992). This argument hopes to avoid the above identified *randomness* or *luck objections* to classical libertarianism – i.e. that injecting the decision or action process with an element of indeterminism is only apt to reduce our freedom because the element of chance induced is outside of our control ("luck decides") –, without having to rely on a mystery or other "panicky metaphysics" (Strawson 1974) to attain its goal of accounting for free will. In essence, agent causation accounts of free will posit a kind of causation that is not part of a causal chain, but rather always the start of a (new) causal chain, thus distinguishing it from regular "event causation" where every cause is also the effect of a still earlier cause. This appeal to a different *kind* of causation is supported by the argument that an agent is not an event, but rather an enduring substance or entity, and therefore not something that can be caused in the manner that an event can be caused (Chisholm 2003). Agent causation ensures free will because choices or actions are neither determined by antecedent events, nor arbitrary or random, but simply caused by the (uncaused) agent (Kane 2005, pp. 44ff).

Agent causation is not without its critics, to put it mildly. The most obvious objection is that "agent causation" is just another name for "mystery", and that nothing has been gained by positing it. This

can also be connected with Strawson's Basic Argument (as mentioned above) to form the claim that even if one allows agent causation, this still does not get us free will: either an agent-caused choice is completely random, or it is made (albeit non-deterministically) on the basis of what might broadly be described as character, and unless the agent is also responsible for forming her own character, she is not ultimately responsible even for her agent-caused choices. This latter objection can, however, be met with appeal to something like Robert Kane's own suggestion of "self forming actions" (SFAs); i.e. those (rare) occasions where a choice between two alternatives is both open (requiring alternate possibilities) and significant for the kind of person we become. Given that the agent herself makes a difference in deciding between the possible options in a SFA, she is also responsible for forming her character, and thus also for the kinds of action she will be presented with later in life. "Ultimate responsibility" (UR) for a given action today can therefore be traced back to such SFAs (Kane 2005, pp. 130-1, 172-3).

Critics have, however, not been placated by the limitation of agent causation to SFAs. Daniel Dennett has objected that given the limitation placed on SFAs, it seems that a person might very well live their life without ever performing one. If this were the case, that person would be without UR, but would, presumably, still feel and appear as free and responsible as anyone (Dennett 2003). Furthermore, one can also counter that, unless one can give an explanation of how agent causation is not just a matter of randomness or luck, SFAs are also ultimately random, and thus not suitable as basis for UR in the sense required by free will.

Given an interest in accounting for free will in face of the challenge presented by determinism, libertarianism does not appear to be a viable solution. In light of this, it is time to look at a third common response to determinism, namely the thesis that determinism (and/or plausible forms of indeterminism) is compatible with free will. This general position is known as *compatibilism*.

4.3 Compatibilism

Compatibilism is a response to the incompatibilist claim that the truth of determinism (and/or plausible forms of indeterminism) necessarily entails that the idea that we have free will is false. Compatibilists are sometimes referred to as soft determinists because many compatibilists accept (as likely) the truth of determinism, but argue that "free will" is possible nonetheless. In support of this they provide various arguments to the effect that determinism leaves room for the kind of free will "worth having", as Daniel Dennett puts it (2003). In all generality, compatibilist responses to the

challenge from determinism engage the concept of free will itself, and try to provide explanations of what really matters to us when we think about freedom and responsibility. Compatibilists have argued, i.a. that what we care about is our ability to act in accordance with second or higher order desires, i.e. desires about what should matter to us. Related to this, some compatibilist frame will and responsibility in terms of having our actions express those values with which we most identify (Frankfurt 1988). According to this picture, we are acting freely when we are able to act in accordance with what we reflectively hold to be our values, and comparatively constrained when we are somehow restricted from acting thus, e.g. by coercion or compelling force. This is an important point, for as shown above, the step from determinism to the thought that our actions are compelled or coerced is easily taken by unquestioned intuition. Nonetheless, as also shown above, determinism entails none of these standard excusing conditions; since this is the kind of thing that matters to our ascriptions of responsibility, determinism is not a threat to our practice of holding people responsible for their actions. Whether or not we can speak of our choices or desires as being uncaused is not merely of secondary importance, it is the wrong kind of question: we *want* our choices to reflect who we are and what we value, and that necessarily includes our history and environment. As already shown, determinism does not entail that our choices and deliberations are ineffectual or epiphenomenal, quite the contrary. Thus, we should not be dismayed by the claim that these also are determined: to all intents and purposes, this only means that they do not arise out of thin air by chance. In other words, determinism can be seen to guarantee the continuity over time essential for any genuine conception of rationality and agenthood (Levy 2007).

Compatibilism has much to commend it, and arguably accords well with many of the things we would pick out when trying to explain why our freedom matters to us. Still, some might think it lacking as an account of freedom in its most fundamental in-the-moment sense. For a long time, compatibilism was taken as the received view, something that has come under attack from new incompatibilist arguments in later years. This is, at least in part, due to the way compatibilists traditionally answer the challenge from determinism, what Harris calls a "bait and switch" (2012): In order to save freedom, compatibilists define it in a way that can also be seen to give up the most important aspects of the concept, namely the intuition that *real freedom must include the ability to do otherwise*. As sketched out above, the compatibilist position does not really answer the incompatibilist allegation that determinism precludes "doing otherwise", and while there have been attempts at reconciling the ability to do otherwise with a deterministic (or plausibly indeterministic) picture by way of the so-called "conditional analysis" of "could have done otherwise", these are

generally taken to have failed, roughly because the conditional "could" opens up the gates for any statements about what could have been done, regardless of human ability (McKenna 2009).

There is another way to challenge the alleged incompatibility of determinism and free will. Considered as a question of justification – does the truth of determinism bear on the grounds on which we justify our beliefs about freedom and responsibility? – the answer, most commonly, has been 'yes'. This is usually thought to be either because those beliefs appear to fall foul of any reasonable metaphysical worldview (e.g. as entailing some form of Cartesian dualism, magical powers, or mysterious agent causation), or because the concept itself is incoherent (e.g. requiring that we both be and not be the causes of our behaviour). Others, most notably P. F. Strawson, have countered that it is wrongheaded to search for justification of those/these beliefs and practices *outside* the practices themselves (Strawson 1974). Taken together with the above-mentioned (apparent) paradox that results from discussing practical consequences of (hard) determinism, there intuitively seems to be something right about this latter approach, as it would appear that the (causal) effectiveness of our moral beliefs and practices insulates them from the accusation that they are illusory. But Strawson's argument is controversial, and though influential, can be considered somewhat of an outlier in the debate. Moreover, it still does not provide an answer to the allegation that determinism excludes an ability to do otherwise – the lack of which will no doubt leave many readers unsatisfied with any solution proffered. I will therefore first spend some time critically investigating the actual entailments of a nomological determinism, as well as presenting a second, radical compatibilist alternative due to Carl Hoefer – which purports to get us this "ability to do otherwise" within the physicalist framework assumed by determinism – before I return to Strawson's fourth alternative in the last section.

4.4 Causal Determinism and Time

So far, I have spoken of determinism both as a matter of deterministic laws and of preceding causes. While the two interpretations in some cases can be used interchangeably, there are also important differences in the conceptual baggage each of them brings with it. I have already mentioned the *nomological* definition of determinism, and under the nomological interpretation, determinism can be understood as the *conjunction* (here effectively the logical statement that two things are true: '&') of a state of the universe (A) at a given time (t_1) and the natural laws (L) which govern the transformation of this state into other states at other times, *entailing* the state of the universe at some other time (e.g. B at t_2). Moving from t_1 to t_2, the state of the universe at t_2, B, is entailed by A&L (Van Inwagen 1975). This is what we say when we say that events are determined by antecedent states of affairs and the laws of physics.

Under the causal interpretation, determinism is often understood as the claim that every event has causes preceding it sufficient for its coming about, what is also known as *causal completeness* (Hoefer 2010). Taken at this level of generality, the distinction between causal and nomological formulations of determinism appears merely terminological. However, the causal interpretation is troublesome in a way the nomological formulation of determinism is not. One reason is that the concept of causality, while extremely common in everyday language and scientific practice, is very hard to precisely define or justify using the kind of physical laws on which determinism is predicated (see e.g. Bertrand Russell's classic put-down in Russell 1912, see also Dowe 2008, Schaffer 2008), and it is useful to be able to talk about determinism while remaining agnostic about the fundamental nature or truth of the separate concept of cause or causality. Another problem with the talk of causality in determinism is the notion of "sufficient cause". While this might seem innocuous at first blush, merely requiring the specification of a set of antecedent conditions which suffice for bringing about the effect in question, closer inspection reveals that such a set would need to be indeterminately large. This can easily be seen if we consider all the negative statements that would have to be included in such a set: my pressing the 'c' key causes a 'c' character to appear in the Word document, *unless* i) the program has hanged, ii) the button is malfunctioning, iii) the battery dies, iv) an asteroid hits the earth, etc. (Hoefer 2002, p. 8). This means that, if we want to posit causal determinism, we need to do so at a universal level, i.e. talk of (entire) preceding states of the universe as causing (entire) following states (Russell 1912, Hitchcock 2007, Hoefer 2002), and given the limitations on the concept of causality as a description of the transformations governed by physical laws, this is effectively just a reformulation of nomological determinism. Anything perceived to be added by using causal talk will therefore have been imported illicitly into the interpretation of that term. The distinction would still be merely terminological but for the fact that two important notions often seen in arguments for incompatibilism appear to be imported in this manner: *direction of time* and *fixity of the past*.

Ordinary talk of causality always implies a specific direction of time, namely going "forward", from past to present or future.[17] In other words, causality is tightly interwoven with our common-sense understanding of a time that *passes*, what J. Ellis McTaggart called the "A-series" concept of time (McTaggart 1908). When we say that something lies in the past, other things are present (however fleetingly), and yet other things lie in the future, we assume the A-series perspective on time. We live inside A-series time, and it is hard to imagine the world in any other light. And yet, the

[17] I venture to use the absolute term here because while there is talk of "backwards causation", this is obviously equally time directed, and only opposed in direction to the regular kind. Note also that backwards causation is a highly controversial subject. See (Faye 2010).

fundamental physical laws on which determinism is predicated are *time symmetric*, i.e. make no distinction between transformations going "backward" and "forward" in time (Hoefer 2010).[18] In other words, when the claim of determinism is made that the conjunction of the state of the universe A at a time t_1, together with the natural laws L, determines the state of the universe B at a time t_2, there is nothing in the laws (L) of the conjunction (A&L) which supports the additional clause that t_1 must be in the "past" – as in "yesterday", or "before his birth" – relative to t_2. Indeed, the laws (of physics) on which determinism is founded can make do with an entirely *tense-less* time series; the "B-series", where different states are specified only in relation to each other as 'preceding' or 'following' (McTaggart 1908, Hoefer 2002). B-series time is static: while things may change from one point to the next, time does not "go by". Consequently, there is no unique, privileged "now", and nothing corresponds to our ordinary concept of "past" or "future". One may still specify a direction of time to describe phenomena as we perceive them (or think about how they happened in the past), but, importantly, determination runs in both (and all) directions[19], and neither discriminates nor privileges any point in space or time as that which determines the rest.

To see the importance of an implied direction of time for our understanding of causality in relation to determinism and freedom, just think what nonsense it would be to say that your choice today *causes* what happened yesterday – if, for example, I were to say that my baking apple cake today caused me to buy apples yesterday. The past is fixed for us in this sense: what has happened is irrevocable. The past – in its "nature", if not retelling – is outside our influence today, in the present. If I want to make apple cake tomorrow, I had better buy apples today. A (technical) way of stating this is to say that the past in A-series time enjoys a special *ontological* status (it "has existed"). What is special about it is that it excludes causal influences going into it, so to speak, while it may be seen as the cause of what is happening now.[20] When determinism is thought of in causal terms, determination appears only to run in one direction, from past to future: "sufficient antecedent causes" becomes sufficient conditions in the past, often the distant past ("before you were born") for added effect. Neither of these are implied by the conjunction of a state of the universe with the laws

[18]While it has long been known that certain forces violate the symmetry implied by relativity (the Charge, Parity, Time-symmetries), the first evidence of time-asymmetry in particle physics comes from recent experiments showing that the oscillations of entangled B mesons between different states have slightly different probabilities depending on which direction the change occurs in. (Francis 2012) While this does indeed show an asymmetry sensitive to the direction of time, it's not quite what we expect or would need to argue that our experience of time in macro-physics similarly constrains micro-physics: there is only a slight difference in probabilities, which, if some frivolity is permitted, is like admitting the slightly lower probability of my baking an apple cake causing me to have bought apples yesterday, than what would be the case if I were to do it the other way around.

[19] This is called the principle of bi-directionality. Note that a weaker form of determinism in which only "past → future" determination is allowed actually poses a weaker challenge to free will than full-on bi-directional determinism. For an explanation of this, see (Hoefer 2002, pp. 6-7).

[20] At least as long as one puts aside the question of how something no longer in existence can have causal influences, but that is another discussion entirely.

of nature (A&L), and therefore not implied by determinism. Although this does not change the fact that determinism excludes the ability to do otherwise, it does undermine one important source of intuitive support for the claim that determinism excludes free will, namely the intuition that we today are (exclusively) determined by events long in the past, i.e. by the Big Bang. The significance of this will be developed below in subsection 4.4.1.

Still, saying that my eating apple cake tomorrow *causes* me to make apple cake today seems obviously false. And while it very likely *is* false (or rather nonsense when considering the meaning of the words used), the falsity or nonsense of it has nothing to do with truth or falsity of determinism (apart from the fact that our ordinary concept of causation might have to be differently specified depending on the truth or falsity of determinism, Hoefer 2010). Rather, our notion of causality and our experience of time – including the experience that the past is fixed and completely outside of our (causal) control, while the present and the future is the battleground of any freedom of will or otherwise – probably have to do with any of the number of *tendencies* observable in nature that are distinctly *uni*directional in time. One popular contender for the bearer of time directionality is *thermodynamics*, the second "law" of which states that systems always tend towards greater *entropy* or elimination of regularity. This is what you observe when you put ice in your drink, and the ice melts, cooling it down. In this scenario, the water in the ice-cubes go from a low-entropy state of being frozen (in which the water molecules are arranged in a regular pattern) to a higher-entropy state of liquid water (in which all the molecules of the now pitifully diluted drink roam around chaotically), equalising in the process the temperature (the average velocity of the molecules) of the surrounding drink and the melting ice cubes (cooling your drink down). The reverse does not *tend* to happen: your drink rarely heats up while the ice cubes grow bigger. While this tendency is universal[21], it does not have the status of a natural law simply because it is not a law, but a statistical regularity. The micro level physics involved are still time-symmetric. You can visualise this if you imagine a speeded-up video recording of ice melting in a drink. Just by looking at this recording, you would immediately be able to tell whether it was being played backwards or forwards. Now, if you were presented with a similar recording of just one (or a small group) of the molecular interactions of the melting ice, you would not be able to say whether the video was being played backwards or forwards; it would be like watching the collision of billiard balls isolated from the context of the cue ball and the players.

[21] As long as one considers a system with enough possible states to make statistical analysis accurate, that is.

Thermodynamic "directedness" arises from the simple fact that there are innumerably more ways for the particles of a system to be distributed uniformly than to be distributed non-uniformly, so that any unconstrained change will tend towards one of these less ordered, more uniform states. Think of billiard balls again: there are innumerable ways the balls can be distributed about the table, and only a vanishingly small amount of them would appear to us as ordered. All the balls sitting in the one half of the table arranged as a neat triangle is just one possible state among a near-infinite number, with the distinguishing feature that it is highly ordered – i.e. that the relative positioning of the billiard balls are far from the average. This, approximately, is the equivalent of a low-entropy state. Any movement of the table or balls would change their ordering into one of the other possible states, progressively moving them away from one we perceive as ordered; it would become disordered, i.e. change towards higher entropy. We are ourselves of course unimaginably complex systems dependent on thermodynamic regularities, and owe our existence to the rise of less complex systems in which the coupling of still les complex systems with opposing tendencies counteracts the dissipation that characterises all transformations at this level (Deacon 2012, pp. 106ff.). For this and other reasons related to the kind of beings humans are, it is highly unlikely that we have the power to alter the past at the level where we perceive events and which is the domain of time-directed systems such as thermodynamics.[22] In other words, McKenna is very likely right in saying that we cannot secure an ability to do otherwise by positing an ability to change the past, at least not at this level (2009). As we shall see presently, things are not so clear-cut if we consider the microphysical level.

4.4.1 Freedom from Determinism

One of the main intuitions behind the incompatibilist argument is the thought that if events in the past determine events now, there is no way for us *now* to act otherwise. As McKenna notes, the only two genuine possibilities of acting differently seems to be if we, through our actions in the present, could either a) somehow violate physical laws, or b) change the past; both of which he quite naturally considers miraculous (2009). Leaving aside for the moment the possibility of violating natural/physical laws and the (arguably failed) conditional compatibilist argument mentioned above, I would like to revisit the assumption against "changing the past" in light of the counterintuitive features of physical laws with which we have now become acquainted.

In a paper entitled *Freedom from the Inside Out* (2002), Carl Hoefer provides an argument to the effect that physical laws can support a notion of freedom based on a non-miraculous way of

[22] We can of course enact great change, but only by work, which itself is a thermodynamic phenomenon, and therefore also directed in time.

"changing" – i.e. determining – the (immediate) past. As already elaborated in the argument above, an important part of the perceived threat of determinism comes not from nomological determinism itself, but from the assumption that determinism is equivalent with the claim that all present action is determined (i.e. *caused*) by past events. In Hoefer's own words:

> The notion of past events determining and explaining future events, and the opposite direction (or an "inside-out" direction) of explanation being somehow wrong or suspect, arises completely from an unholy marriage of A-series time with deterministic physics. (2002, p. 5)

Taking the four-dimensional "block universe" view of space and time, our present can be considered a time-slice somewhere in the middle of that block. Because it is not inherent in the model or the laws on which it is based to differentiate between any possible time-slice, *we are free to posit the present time-slice as determining the rest*. As already noted, the usual worry is that such a statement is tantamount to positing backwards causation, something which very likely *is* impossible (at least for humans) independently of whether we assume A- or B-series time (Hoefer 2002, p. 5). To reiterate, this worry does not arise from determinism itself, but rather from the additional (and arguably highly problematic) causal interpretation of determinism. Forget the troubling thought that we should somehow be able to *cause* the past, and get used to the idea that determinism is compatible with regarding the present as *determining* the past. Hoefer argues that this "determinism from the inside-out" can provide a notion of freedom acceptable both physically and for common sense (2002, p. 6).

There are two central points to Hoefer's argument. The first point is the one outlined above, namely that we are free to consider determinism "from the inside out", i.e. from the time-slice of block space we consider "present" outwards into the two halves we usually think of as past and future. From this perspective, we are free to say that our actions in the present determine the rest, and therefore past and future states. It remains true that everything in the present time-slice is determined, as is everything wherever and whenever if we assume the truth of determinism, but as Hoefer puts it, our actions are "simply determined"; meaning not unfree or caused or anything else which one usually concludes from the "unholy marriage" of determinism and A-series time (2002, p. 4). The crux of this is that we can regard our actions as *determining* rather than *being determined* (by something else).

It is this radical change in perspective that does most of the work in Hoefer's argument, but there remains a large conceptual hurdle before this perspective can deliver on the requirement of an ability to do otherwise: Even taking the inside-out perspective, it seems that the past, *being as it is known to*

us, still necessitates events in the present. To answer this objection, Hoefer again appeals to the difference between the past at the macro-physical level at which we are familiar with it, and the micro-physical states which underpin these. Hoefer's second point is based on the claim that while an "inside-out" determining action in the present does affect past and future, it does so only by placing *constraints* on possible past and future states. Keep in mind that neither past nor future are *ontologically* special in this framework, so placing a constraint on preceding states is not beyond the pale. Accepting all the (macro level) constraints implied by an assumed context (I am sitting at a table before my laptop, etc.), the constraints imposed by my action now also turn out to be so weak as to be *practically* negligible for actions with a limited extension in space-time, such as typing on a keyboard.[23] The reason for this is two-fold: 1) for any given macroscopic event (like that of typing 't'), there will be innumerable micro-physical states that could correspond to that event (think only of all the different configurations of molecule placement and velocity that can correspond to the average temperature in the vicinity of my typing),[24] and 2) any roughly delineated macro level event – say with a 10 m radius – will only entail constraints at the level of micro-physical states on a "time-cone" in each time-direction with a radius decreasing to 0 after approximately 3.3×10^{-9} seconds (the time it takes light to travel 10 m). In other words, my typing 't' now will only determine space-time at a micro-physical level in a very limited (and vanishing) area for an extremely short time. Practically nothing is entailed about the preceding state of affairs at the macro level of events that are recognisable by us. The only requirement is that the preceding events at that level must have within their "supervenience base" a micro-physical state compatible with my typing 't' in the following moment.[25] Now for the freedom: given a known past, i.e. given a known series or

---

[23] Typing on a keyboard can of course have tremendous effects in the future, e.g. deciding whether someone graduates or not, and so can be seen to place (important) constraints on *identifiable* future states however far into the future you would care to follow its trail. But keep in mind that the constraints introduced by pressing a single key at a given time now will wash out almost completely when taken together with the incomprehensibly large number of events taking place at any given point in time also placing constraints on the future. My typing 't' now is in no way sufficient for bringing about me graduating on its own, just as Lorentz' proverbial butterfly is but an infinitesimal part of the processes leading to hurricanes in Texas. Thus the real effects of our actions that we can perceive are the results not of single free actions in the present time-slice determining the rest, but rather a complex phenomenon on the level of macro-physics (and higher levels of organisation/regularity) for which we have no problem of accounting in terms of regular cause and effect, regardless of any talk of micro-level determinism, and, importantly, perfectly compatible with regarding that choice as free. This is true the "other way around" as well, namely if we think of the constraints imposed on our freedom now given the past as this is known to anyone. No one is trying to deny these perfectly obvious constraints: I would not be able to type freely on the keyboard of this laptop had I not brought it with me to the library, etc. – but that was also a free action within unproblematic macro-level constraints.

[24] Hoefer assumes that identifiable events can be considered event types at the macro level and that these can have many instantiations on the micro-physical level: "We assume, in other words, that there is some ill-defined and probably infinite set of microphysical state-types that are "good enough" to count as a supervenience base for my typing "t" in the assumed context."(Hoefer 2002, p. 7).

[25] Note that while I still seem to be wedded to A-series time here, this is in fact only a device applied to help make visible the kind of constraints involved. I could equally well say that the events following this typing must have within their supervenience base a micro-physical state compatible with me typing 't' now. But due to the fact that our perspective *is* from inside A-series time, this future-directed constraint-formation appears to us unproblematic, and it is necessary to distinguish between the causal influence our deterministic actions have on the future from the deterministic "influence" our actions (according to the inside-out perspective) have on states preceding and following that action.

set of events recognisable at the macro-physical level in the period of time preceding the present, I am now, in accordance with the rules of grammar, free to type 's' or 'z' when I write 'recognisable'/'recognizable' (or 's' *and* 'z', like I just did), and whichever I choose to type determines the preceding and following micro-level states in a vanishingly small area around me for an incredibly short time.

There is now nothing standing in the way of regarding that choice as perfectly free in the sense we are after here. I really could have typed only 'z' or 's', or relied on another example altogether. Of course, the example is adapted from Hoefer, and so I was inclined to use just those letters from reading his article – still, I was not determined by the Big Bang (or whatever lies at the start of everything) to write it in just that way. While it remains true within this picture that my choice of what letter to type was deterministic, the physical picture entailed by that adjective is inherently perspectiveless and does not support any inference to the status of that choice as free or unfree. In order to understand anything from this metaphysical notion, we have to give it a perspective and an interpretation. "Freedom from the inside out" can thus be seen as the application of a particular perspective, a perspective that arguably accords with the (commonsense) requirement that freedom entails an ability to do otherwise and which can supplant the standard perspective in every way relevant to the thesis of determinism.[26]

Hoefer's argument is perhaps counterintuitive, and it admittedly depends on refutable assumptions, but so does the standard argument for incompatibilism outlined above, and to the eyes of anyone who would lament the loss of free will, this also suffers from other serious defects. [27]

However, my aim here is actually not to present a coherent picture of freedom within the available deterministic framework (even though I think Hoefer's is an interesting one). Rather, I am trying to undermine the traditional incompatibilist picture where the truth of determinism entails the falseness of freedom (in one intuitively important sense of the word). To the extent that I have succeeded in undermining the "ability to do otherwise"-argument of incompatibilism, what am I left with? While I think I have made difficult the "obvious" conclusion that freedom is impossible in physical terms, I certainly have not provided a watertight argument for the compatibility of freedom and determinism (or indeterminism). In truth, I am left with a negation, but one I hope to show is

[26] Hoefer also shows how it is compatible with (physically plausible accounts of) indeterminism (see Hoefer 2002, pp. 12-4).
[27] Notably, both rely on assumptions about physicalism/materialism, supervenience and reductionism. Standard incompatibilist arguments additionally rely on an interpretation of time not entailed by the physics involved in their claim. I also personally count against standard incompatibilism the practical inconsistency pointed out above (in section 4.1.1) concerning the way we should relate to its claim that we are unfree and therefore essentially not responsible for our choices of what to relate to.

as good as any compatibilist account in diffusing the worry that genuine freedom is somehow ([meta-] physically) impossible.

Having thus surveyed the debate of what van Inwagen calls the "mystery of free will" (Inwagen 2000), I am inclined, but not hard-determined, to take the next step in a different direction.

5 Leaving Behind the Metaphysical Challenge

> "[A]rgument, reasonings, either for or against the sceptical position, are, in practice, equally inefficacious and idle […]. [… W]e have an original non-rational commitment which sets the bounds within which, or the stage upon which, reason can effectively operate, and within which the question of rationality or irrationality, justification or lack of justification, of this or that particular judgement or belief can come up." P. F. Strawson (2005)

If nothing else, the preceding sections should have left you with a sense of just how complicated the question of the (in-) compatibility of freedom and determinism really is. Far from being resolved to any claim's favour, there are fundamental issues with each of the three major positions that have no apparent resolution beyond a reliance on the relative strength of the intuitions which support each argument; whether you lean towards hard determinism (or impossibilism), libertarianism or compatibilism, your choice ultimately relies on preference.

I have already spoken of determinism as presenting a "challenge to common sense" in that it appears to conflict with notions of freedom and responsibility that are both widespread and widely believed true. Provided that this metaphysically based challenge leaves us with no clear answer regarding the reality of free will and moral responsibility, it is an arguably reasonable next step to investigate what other ways there might be to frame these dual notions that seem to be so important to the way we see ourselves and lead our lives. If we can give an alternative, plausible grounding for these ideas – and by grounding I mean something beyond a mere appeal to consequences, i.e. beyond an appeal to the detrimental effects the acceptance of the truth of incompatibilism would have (or lack thereof, cp. Roskies 2006a); I mean a good reason to keep on seeing ourselves as free and responsible that holds independently of any practical reason we might have for keeping our beliefs, since such a direct appeal to practical consequences tends to leave a bad taste in the mouth of anyone sufficiently worried by the possibility of incompatibilism to seek out an answer to the question of whether free will is possible – if, then, we could give an alternative grounding for our beliefs, the lack of a resolution to our original question might fade in importance sufficiently to be left behind without much discomfort.

5.1 Naturalism: A Fourth Option

The debate concerning free will and determinism as we have surveyed it here is at an impasse, and has been for a long time. New arguments are presented on each side with some regularity, but are equally regularly found lacking by the opposing side. Is there any possibility for a reconciliation? In

a seminal essay entitled *Freedom and Resentment* (FR, 1974)[28], P.F. Strawson attempted such a reconciliation. While Strawson's approach is typically considered a compatibilist argument in the classical sense, the unique way he went about making it means that his response can also serve as a "fourth option" to the three standard positions already considered, effectively allowing us to lay what arguably is a sceptical challenge to rest. In this light, Strawson's argument falls into the second category of responses to the challenge from determinism, what I at the start of Chapter 4 identified as rejecting the challenge from determinism, rather than attempting to answer it. In order to see the potential of this fourth option, we must first consider Strawson's view of the classical debate as laid out in *FR*.

5.1.1 The Optimist and the Pessimist

Strawson sets up an argument between a pessimist and an optimist, characters representing incompatibilism and compatibilism, respectively, whose names point to the fact that the discussion is not only theoretical, but has practical and emotional impact on our lives. The pessimist is therefore not merely pessimistic about the prospect of giving a coherent account of freedom and responsibility compatible with the truth of determinism; he palpably depressed or dispirited at the thought that belief in free will appears unfounded (McKenna and Russell 2008, p. 4). The optimist, as we might expect, tries to convince the pessimist that we can find sufficient justification for our moral practices by other means, i.e. by appeal to consequentialist reasons – such as prevention in the case of punishment –, and that nothing of importance is lost to the truth of determinism. The pessimist remains unconvinced, feeling that the optimist is leaving our something crucial, some deeper sense of what it means to be free and responsible. Strawson accepts the pessimist's criticism of the optimist, but rejects the standard (non-)solution, proffered by libertarians, of denying determinism and attempting to account for some "extra factor" that guarantees freedom; what Strawson, as mentioned above, calls "panicky metaphysics".

5.1.2 Our Reactive Attitudes and when We Suspend Them

In order to attempt a reconciliation, Strawson asks us first to take a step aside from the words and concepts we usually apply in this discussion, things like "agent causation", "conditional arguments" etc., and train our gazes on something closer to home, namely on

> […] the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or

[28] Read in (McKenna and Russell 2008, pp. 19ff). I follow here the practice of McKenna's reprint of Strawson's text which refers with bracketed numbers [60] to the page number in the version of Strawson's text published in (Watson 1982, pp. 59-80).

involve, our beliefs about these attitudes and intentions. (FR [62], McKenna and Russell 2008, p. 22)

McKenna describes this as "Strawson's 'naturalistic turn'", turning away from the usual conceptual analysis and looking at what goes on when (e.g.) we *hold* someone responsible (2008, p. 5). This, warns Strawson, necessarily involves a loss of precision, since the field of everyday activity in which we are now moving is so complex. But the loss of precision is arguably made up for by the gain in validity that comes from discussing the phenomena as they present themselves (in all their complexity), rather than easily definable concepts suitable (or made to suit) for a conceptual analysis.[29]

Strawson wants us to focus on an important subset of our personal feelings and reactions, namely those that depend on our beliefs about the feelings and attitudes of *others* towards ourselves. He calls these our "reactive attitudes". If someone steps on your hand, what you believe about the attitudes and intentions expressed by that someone towards yourself in the act of stepping on your hand has tremendous importance to how you react. If you think that he knew your hand was there, but stepped on it anyway, you would be (rightfully) indignant from his lack of regard. If, additionally, you (reasonably) believed that he stepped on your hand with the express intention to cause you harm, righteous anger would (arguably) be in order. And so on for a vast range of reactive attitudes concerning the ill-or-good-willed actions of others towards yourself, as well as your reactive attitudes towards your own actions, e.g. the guilt you might (correctly) feel at having wrongfully disregarded the interests of a fellow.

Having identified this commonplace feature of human interaction as our present field of study, Strawson asks us to consider which situational features cause us to suspend these otherwise pervasive reactive attitudes. He identifies two categories of reasons that he thinks are sufficient for his purposes: 1) specific situations in which we temporarily suspend our reactive attitudes for one reason or another, e.g. for those reasons that make it appropriate to say "he didn't mean to", "he hadn't realised", "he couldn't help it", "he had no alternative", etc. What is common for all the various instances where one of these might apply is that we do not suspend our reactive attitude towards the person acting, but rather towards this single act, the specific injury inflicted. In other words: "They do not invite us to see the *agent* as other than a fully responsible agent. They invite us

[29] This kind of "pragmatic" approach to philosophy that trades formal accuracy for descriptive power has many detractors, but keep in mind that we are exploring Strawson's argument as an alternative to the formally more stringent, but woefully deadlocked traditional approaches.

to see the *injury* as one for which he was not fully, or at all, responsible."(FR [65], McKenna and Russell 2008, p. 23).

The second category has two subcategories, the first of which contains the instances in which it is appropriate to temporarily suspend our reactive attitudes towards *an agent*, e.g. the times at which it can be appropriate to say things like "he isn't himself", or "he is under great duress at the moment". While these pleas have their place, they are far less important (for present purposes) than the second subcategory. To this belongs the cases in which we *permanently* suspend our reactive attitudes towards an agent (and therefore towards his actions). The *circumstances* are in this case normal – the agent "was himself", so to speak – but *he* is "warped or deranged, neurotic or just a child."(FR [66], McKenna and Russell 2008, p. 24).

5.1.3 The Participant and Objective Attitudes

From this, Strawson draws up a "crude dichotomy" between two ways of relating to other humans, a "participant attitude" and an "objective attitude". When we suspend (all) our reactive attitudes towards an agent for reasons such as those which belong to the second subcategory above, we are taking an objective attitude towards that person, seeing him "[…] as a subject for what, in a wide range of sense, might be called treatment[…]" (FR [66], McKenna and Russell 2008, p. 25). Taking an objective attitude does not mean suspending all emotional responses: one may be repulsed, say, or pity the person towards which one adopts an objective attitude, but the range of emotions (or cognitive-emotional complexes) expressed in terms of reactive attitudes are excluded from the objective relation. By way of example, in my work as a nursing assistant for seniors with dementia I may sometimes appropriately react with anger when one of the patients that I care for is violent, but I may not feel resentment towards them, at least to the extent that I have (appropriately) adopted the objective attitude of a nurse caring for a patient no longer the master of themselves.

Strawson submits that we *can* take this objective attitude also in "normal" cases where none of the above categories applies, i.e. regard all our fellows and interpersonal reaction with an analytical eye from an impersonal standpoint. Still, being human, we are not able to keep this up for long: to suspend (all) our reactive attitudes is akin to withdrawing from life itself, at least to the extent that this entails meaningful and valuable interpersonal relations (FR [67], McKenna and Russell 2008, p. 26). This is similar to the problem that faces anyone trying to maintain a radical, sceptical doubt with regard to knowledge: however hard you try to maintain a pervasive doubt, you will soon

succumb, by the nature of practical life above all, to the need to assume that some things are some ways and others different.

Furthermore, that the standard modus of humans and human interaction is one of participant attitude is not a claim that needs to be justified or defended by anything but reference to the phenomena to which this term is defined to refer: "This is simply a 'given' of our human nature and without it we would hardly recognize an individual or community as being fully *human*."(McKenna and Russell 2008, p. 6, emphasis in original)

Given these distinctions and the conditions in which it is appropriate to suspend our (otherwise pervasive) reactive attitudes towards actions or agents, we can now ask what influence the truth of the thesis of determinism could have on these attitudes, and whether there is any way in which this truth should lead us to suspend them.

5.1.4 Why Determinism does not License the Objective Attitude

Strawson answers this question on three levels, beginning with the question of whether the truth of determinism would bring into action any of the two categories in which it is appropriate to suspend our reactive attitudes towards actions or agents. This is answered in the negative: the truth of determinism neither makes it so that no action is done with good or ill will, nor makes it so that everyone is under duress to do what they do, nor activate any of the other reasons we have for suspending our reactive attitudes towards the actions of others. Nor does the truth of determinism make it true that people, in general, are psychologically abnormal or children (a self-refuting hypothesis), so that we should adopt the objective attitude towards other people (and ourselves) in general.[30]

Beyond this, even if there were some way for the truth of determinism to make it appropriate pervasively to adopt the objective attitude, humans would be *incapable* of doing this. However hard we may try, convinced by some traditional hard-determinist argument, we would be forced by our very nature into treating most people (and ourselves) as agents with whom it is appropriate to have participant interaction, (if for no other reason than to distinguish them from those that clearly merit the kind of treatment entailed by adopting an objective attitude, i.e. those that are warped, deranged, neurotic or just children).

[30] This also predicts Stephen Morse's distinction between causation and abnormal causation in the actually non-analogous cases of a person behaving abnormally due to e.g. a brain cancer, and a person's behaviour being the result of normal brain processes – and analogy often, and falsely, adduced in arguments from neuroscience and determinism to hard determinism. See (Morse 2010).

Finally, even if we *were* given the "god-like" choice of whether or not to suspend all reactive attitudes, this would mean choosing between an emotionally impoverished, dehumanized life and its opposite; a choice on which the truth of determinism could not rationally be claimed to have any say.

As for the typical "moral judgements" of right and wrong, Strawson says of them that they are generalised, "vicarious analogues" of reactive attitudes. This means, roughly (and amongst other things), that when you see someone treating me in a certain way, *you* may judge, based on features of the situation, that *I* am justified in being indignant: the treatment can be considered mistreatment, or as displaying ill will, and therefore can be judged morally wrong. Therefore, the same applies to moral reactive attitudes and judgements as to personal ones: the truth of the thesis of determinism does not make it appropriate to suspend them; neither in specific cases, nor in general.

5.1.5 Attempting a Reconciliation

From this way of conceiving of the issue, Strawson revisits the argument between the optimist and the pessimist, and points out using the tools now acquired just what he thinks each of them has gotten wrong. The optimist typically argues that the truth of determinism does not affect the efficacy of rewards and punishments in regulating behaviour, e.g. punishing to prevent further crime. But this picture is "wholly dominated by objectivity of attitude", with "[t]he only operative notions involved [being] such as those of policy, treatment, control." By excluding the moral reactive attitudes, the optimist's picture is leaving out "essential elements in the concepts of *moral* condemnation and *moral* responsibility."(FR [76], McKenna and Russell 2008, p. 7, emphasis in original). The pessimist is right to dismiss this attempted solution, but he is wrong in thinking that the reactive attitudes are insufficient to "fill the gap in the optimist's account", and that he must therefore provide some metaphysical explanation that can be "repeatedly verified" in all cases where we want to claim that someone is acting freely and is morally responsible (FR [79], McKenna and Russell 2008, p. 7). Only if the optimist concedes to the pessimist that something vital is missing from his picture, and only if the pessimist in turn lets go of his metaphysics, can the two be reconciled and the dilemma of determinism hope to find a solution. At present, the debate is at a standstill because both parties are guilty of "over-intellectualizing the facts"; disregarding the emotionally rich framework in which these practices actually exists, seeking some purely intellectual, outside justification:

> Inside the general structure or web of human attitudes and feelings of which I have been speaking, there is endless room for modification, redirection, criticism, and justification. But questions of

> justification are internal to the structure or relate to modifications internal to it. The existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external 'rational' justification. Pessimist and optimist alike show themselves, in different ways, unable to accept this. (FR [78-9], McKenna and Russell 2008, pp. 7-8)

McKenna points out two important aspects of Strawson's naturalistic turn that have interesting historic precedents. The first is his emphasis on emotion in moral judgements, what at the time of writing *FR* was a return to the "moral sense theory" associated with David Hume and Adam Smith. This emphasis on emotions has gained prominence in wider circles interested in morality in recent years, especially as the human sciences have picked up the topic, giving rise to subfields such as affective neuroscience. That emotions are not "rationally suspect" – a historically common prejudice in much philosophy –, but rather an essential part of rationality itself, is a realisation with increasing influence in both philosophy and science (see e.g. Nussbaum 2001). The other, for purposes of this section more important aspect of Strawson's rather unique contribution is what McKenna calls Strawson's way of "discrediting the skeptical challenge" that the standard dilemma of determinism can be taken to represent, something he does by reference to "the inescapable, psychological mechanisms that guide human thought and action."(McKenna and Russell 2008, p. 8). As McKenna points out, Strawson develops his ideas from *FR* in the later book *Skepticism and Naturalism* (SN, [1985]/2005), in which the connection between his approach to the debate concerning free will and determinism in *FR* is connected to a kind of "soft naturalism" taken up, i.a. by Hume, in reply to the classical forms of epistemological scepticism, such as scepticism about secure knowledge of an independently existing world.

This is, in my eyes, the crux of Strawson's approach, and why I have presented his argument as a "fourth alternative" instead of reporting it under the more commonly applied heading of compatibilism. For, while one might be tempted to think that Strawson's appeal to the fact that we neither can nor could want to rid ourselves of our reactive attitudes constitutes a concession to the hard determinist/incompatibilist, in that the solution proffered is *merely* pragmatic, leaving the fundamental challenge unanswered and therefore (tacitly) admitted (as in the traditional compatibilist accounts), subsuming Strawson's argument under classical compatibilism in this manner misses the main, if not the entire, force of his argument. To see how Strawson's suggestion might actually reconcile the pessimist and the optimist, we should therefore investigate more closely what Strawson calls "the way of Naturalism"; "a different kind of response to skepticism – a

response which does not so much attempt to meet the challenge as to pass it by." (Strawson 2005, p. 11 & p. 3).

5.1.5.1 Soft Naturalism in Reply to Sceptical Determinism

Strawson covers a much wider ground in *SN* than in *FR*, and so the general outline of his reply to determinism first appears in his comparison of the way Hume and Wittgenstein relate to the justification of judgements:

> [Hume and Wittgenstein] have in common the view that our "beliefs" in the existence of body and, to speak roughly, in the general reliability of induction are not grounded beliefs and at the same time are not open to serious doubt. They are, one might say, outside our critical and rational competence in the sense that they define, or help to define, the area in which that competence is exercised. To attempt to confront the professional skeptical doubt with arguments in support of these beliefs, with rational justifications, is simply to show a total misunderstanding of the role they actually play in our belief-systems. The correct way with the professional skeptical doubt is not to attempt to rebut it with argument, but to point out that it is idle, unreal, a pretense; and then the rebutting arguments will appear as equally idle; the reasons produced in those arguments to justify induction or belief in the existence of body are not, and do not become, *our* reasons for these beliefs; there is no such thing as *the reasons for which we hold* these beliefs. We simply cannot help accepting them as defining the areas within which the questions come up of what beliefs we should rationally hold on such-and-such a matter. (Strawson 2005, p. 21, emphasis in original)

There are two elements at play here. 1) Strawson is saying that "professional skeptical doubt" about the grounds for our belief in the existence of "body" (the things that populate the perceptible world) and in the reliability of induction, is idle, pointless: those "beliefs" are not the kind of beliefs that need to be justified. Rather, they serve as part of the framework *within which* it is possible to justify beliefs proper. Thus, when we in a particular case investigate whether or not we are right about some judgement of what something in the world is, we do so only on the backdrop of the belief that things in the world are a certain way, and that we can make reliable judgements about them. That is not to say that our practices are what validate these beliefs, or that scepticism about these beliefs is self-refuting because talk of doubting them presupposes certain foundational concepts (Strawson 2005, pp. 3ff). No, these beliefs are simply outside of justification, not something we can rationally doubt or reason away. We may try, as the sceptic indeed does, to doubt them and demand justification for why we should believe in them, but try as we might, we cannot escape relying on them in anything and everything connected with human life. Hume saw them as simply given by

Nature; that we cannot any more choose to judge than we can choose to breathe, and Wittgenstein, roughly, saw them as part of the framework or scaffolding of our thoughts.

2) Because of this, trying to argue with the sceptic is equally idle and pointless: these beliefs are also *not in need* of any justification. The framework given by nature is not something we have assented to; is what allows us to distinguish between true and false in the first place (as Strawson quotes Wittgenstein saying, Strawson 2005, p. 16). Thus, if we disregard this and go looking for reasons to continue to hold a belief, e.g. reasons or justification for believing in the reality of the external world, whatever reasons we might find would still not be (or become) *our reasons* for believing in the external world– *we* have no such reasons, we simply have these (framework) beliefs.

That is not to say that science or reasoning is pointless, of course, but that science and reasoning are properly occupied with questions within the framework of some given beliefs.

Strawson argues that this is directly transferrable to the debate concerning free will and determinism, since, as he has analysed our belief in free will, the reactive attitudes (and their vicarious analogues the moral judgements) are just as much a part of our given nature or "humanity" as our belief in physical objects:

> What we have, in our inescapable commitment to these attitudes and feelings, is a natural fact, something as deeply rooted in our natures as our existence as social beings. (Strawson 2005, p. 34)

Our reactive attitudes are, in the end, "inescapable" says Strawson.

Although that "inescapable" must be qualified, seeing as how we do have the ability, in certain instances and for certain people, to suspend our reactive attitudes and take what he called the "objective attitude" towards such people. And as the "moral sceptic" (the incompatibilist) based his argument against our unreasoned views on just this our ability to view each other from a detached, objective point of view, the question can then be raised: from which point of view, the participant or the objective, do we see things as they really are? For, even if we might not be able to escape our reactive attitudes completely in practice, it seems that we can still say of the objective point of view that it is the right one, the one most in accordance with how the world really is (Strawson 2005, pp. 36ff). It seems that we are still stuck with the same problem as before. This, however, is only appearance: we have established the inescapability and unjustifiability of our reactive attitudes, and the existence of a second point of view with a second set of attitudes can now be granted without discomfort. Indeed, the whole problem, continues Strawson, arises, not from the existence of two possible points of views that are, if not mutually exclusive, at least tending to exclusivity "at their

limits"; the problem arises from the forced choice between them. But there is no reason to think that only one is correct or true, or that one is "more true" than the other; they are fundamentally different ways of looking at human action. The real illusion is the idea that there is a "metaphysically superior standpoint":

> Once that illusion is abandoned, the appearance of contradiction is dispelled. We can recognize, in our conception of the real, a reasonable relativity to standpoints that we do know and can occupy. Relative to the standpoint which we normally occupy as social beings, prone to moral and personal reactive attitudes, human actions, or some of them, are morally toned and propertied in the diverse ways signified in our rich vocabulary of moral appraisal. Relative to the detached naturalistic standpoint which we can sometimes occupy, they […] have no properties but those which can be described in the vocabularies of naturalistic analysis and explanation. (Strawson 2005, p. 39-40)

Note Strawson's usage of "naturalistic" here to signify the "sceptical" or objective position. This is where Strawson introduces his distinction between a non-reductive or "soft" naturalism and a reductive, "naturalistic" view of humans. Whereas the latter form of "hard" naturalism seeks reductive explanations of human phenomena in the terms of purely objective natural sciences, Strawson argues that a soft naturalism can incorporate both the scientific objective regard with its absence of interpersonal reactive attitudes, and the personal perspective; and thus, does not try to deny or seek (unneeded and unwanted) objective justification for how we as humans relate to the world, each other, and ourselves.

To substantiate this "relativizing move" (2005, p. 38), Strawson compares the participant and objective attitudes (as complementary, moving on mutually exclusive, points of view) with the case of perception and its objects. For, argues Strawson, we can talk of the objects of perception either as being, in the vulgar meaning of these terms, "coloured or plain, hard or soft, noisy or silent" (2005, p. 52); or we can talk of them in the terms of physical theories as having certain physical properties that impinge on our sensory apparatus, modulating neuronal patterns of excitation, etc. Which of the two descriptions or explanations we are interested in depends on what point of view we inhabit. Are we asking what something looks like, or are we wondering how the visual system works? Of course, neither point of view, whether it is subjective or objective, on human relations or objects of perception, is above error or criticism: speaking of the subjective point of view, we can certainly be mistaken in our perception of something or in our reaction to some perceived harm, and it may be argued that large parts both of perception and our reactive attitudes are shaped by our particular history and current situation, and that the subjective view therefore has "internal relativities". While

this on the one hand can make it even easier to accept that there are no single metaphysically superior point of view of the world as it "really is" (since there is also no singe subjective point of view under which every subject would have to be subsumed), this point can also be exploited by the "scientistic hard-liners" to argue that *their* framework – which, while struggling with internal relativities of its own, has been build up over time using a variety of tools to ensure a cumulative advance in knowledge ("notably, the test of verification, the hard test of success or failure in prediction and control") – is clearly superior to the multifarious and uncertain subjective point of view (2005, p. 51). Moreover, the hard-liners or hard naturalists might further claim that their objective view of the world can account, not only for the world as it really is, but also for every phenomenon described by the rival subjective views, namely *as* subjective experiences and reactions for which the hard naturalist can give a "causal analysis" (Strawson 2005, p. 52).

This, however, misses the point and force of Strawson's "relativizing move", and here a lengthy quote must be excused:

> To this the patient reply must be: not that it is mistaken to think of the physical world in the abstract terms of physical science which allow no place in it for phenomenal qualities; nor that it is mistaken to think of the world of human behaviour in the purely naturalistic terms which exclude moral praise or blame; only that it is mistaken to think of these views of the world as genuinely incompatible with the view of physical things as being, in the most unsophisticated sense of these words, colored or plain, hard or soft, noisy or silent; and with the view of human actions as, sometimes, noble or mean, admirable or despicable, good or evil, right or wrong. Though we can, by an intellectual effort, occupy at times, and for a time, the former pair of standpoints, we cannot give up the latter pair of standpoints. This last is the point on which the non-reductive naturalist, as I have called him, insists. What the relativizing move does is to remove the appearance of incompatibility between members of the two pairs of views. Without the relativizing move, the scientific hardliner, or reductive naturalist, could stick to his line; admitting that we are naturally committed to the human perceptual and morally reactive viewpoints, he could simply conclude that we live most of our lives in a state of unavoidable illusion. The relativizing move averts this (to most) unpalatable conclusion. It would surely be an extreme of self-mortifying intellectual Puritanism which would see in this very fact a reason for rejecting that move. (Strawson 2005, p. 52)

By treating the deterministic dilemma as a sceptical challenge and connecting his reply to established philosophical tradition, Strawson goes a long way in diffusing the entire worry that our beliefs and practices concerning and surrounding free will and moral responsibility are somehow untenable or in need of metaphysical justification.

Interlude

6 Questioning the Framework or Framing our Questions?

> "One reason why most people don't perceive [incompatible beliefs in free will and in determinism] as a conflict might be that our belief in freedom is so deeply embedded in our everyday thoughts and behavior that the rather abstract belief in physical determinism is simply not strong enough to compete. The picture changes, however, with direct scientific demonstrations that our choices are determined by the brain. People are immensely fascinated by scientific experiments that directly expose how our seemingly free decisions are systematically related to prior brain activity."
> John-Dylan Haynes (2011b)

6.1 Free Will as Framework: Taking Stock

Taking a cue from Strawson, most of the discussion in Part I can be seen as a discussion of the *framework* of free will, i.e. the framework constituted by our ideas about freedom and responsibility, *within* which we can distinguish between actions that are free in the relevant sense and those that are not, and between persons that are responsible in the relevant sense and those that are not. When discussing determinism and indeterminism, incompatibilism and compatibilism, we were never concerned with whether some person A, in some situation X, had acted freely, or if it was appropriate to hold person A responsible for doing X, or to what degree we should hold A responsible for X; no, we were considering whether it is *ever* accurate, or even permissible, to talk of a person doing something freely, acting out of their own, free will, or if it is *ever* legitimate to say that someone is responsible for doing something, in the sense of 'responsible' that goes beyond mere physical causation or strict liability.

Conversely, outside of this philosophical scientific discourse, the reality of free will and responsibility is presupposed, taken for granted, forming the framework within which we resolve the questions that matter to us in the present, namely questions about responsibility for specific people in specific cases: Who was responsible for this? Could they have avoided it? What was their intent? Should they have realised the consequences? When we judge our fellows or give them fair trail, we typically do not ask: Is there free will? Is the notion of moral (and legal) responsibility supported by a coherent metaphysics? No, we want to know *who did this*, and whether or not it is right to hold that person responsible for that action and its outcome (Greene and Cohen 2010).[31]

[31] Stephen J. Morse, speaking on the relationship between neuroscientific results and legal practice, calls the first kind of discussion "external critique", as it is external to the framework of our ascriptions of legal responsibility, while the

In surveying what admittedly are only parts of this vast "framework debate", we saw that the question of whether free will is compatible with determinism, or indeterminism, or any coherent metaphysics for that matter, is complex and far from resolved. I went to some lengths to undermine the felt acuity of the incompatibilist dilemma (that free will seems to be impossible whether the world is deterministic or indeterministic) with the help of Hoefer's radical, if counterintuitive, shift in perspective from outside-in "determined" to inside-out "determining".

Finally, considering determinism's "challenge to common sense" a sceptical challenge to the notion of free will and moral responsibility, I argued with P. F. Strawson that it was misguided to search for external justification for our "reactive attitudes" and their vicarious analogues the moral judgements: What Strawson called our "personal and moral reactive attitudes" are given features of human life, not suppositions requiring or capable of being given justification. According to Strawson, all three main positions in the standard debate concerning free will and determinism are guilty of over-intellectualising the issue, accepting that the framework formed by reactive attitudes and moral judgements is in need of rational outside justification. This error is, at least in part, the result of a form of scepticism with strong similarities in structure to the kind of epistemological scepticism that both Hume, Kant and Reid dealt with, a scepticism that, according to what Strawson called "soft" or "non-reductive naturalism", is idle and pointless; perhaps, at best, showing the limits of human reason, but by no means providing reason to *actually* doubt all claims to knowledge, or in the present context, undermining the appropriateness of our reactive attitudes *in toto*.

6.2 Empirical Critique of "Free Will as Framework": The Discussion Ahead

The initial aim of Part I was to separate this philosophical debate about the *possibility* of free will from the kind of critique that neuroscience (along with the other, empirical sciences relevant to human reasoning and behaviour) could leverage against our traditional conception of free will and responsibility. The human sciences do not touch upon the question of whether the world, fundamentally speaking, is deterministic or not, nor whether it is possible to give an account of free will that is compatible with any reasonable metaphysic. Therefore, neuroscience does not have a say in this fundamental debate, and cannot provide or partake in any "wholesale" denial of the framework of our traditional conception of free will and responsibility that would arise from it. Nor can neuroscientists eager to explain (free) will in mechanistic terms rely on the fundamental incompatibility of freedom and determinism (or indeterminism) when claiming they have shown

---

second kind is dubbed "internal critique", since in this case, neuroscientific evidence is applied to settle questions within the framework of the assumption of "minimal rationality" that is operant in (US) courts of law (Morse 2010).

free will to be an illusion, since no such simple incompatibility exists; despite what Haynes appears to believe, his Libet-style experiments are not empirical confirmation of an already-accepted truth about the incompatibility of free will and determinism.

Neuroscience, then, has no truck with the fundamental philosophical debate concerning free will and determinism.[32] That being said, one might still make a scientific critique of the framework of free will. Instead of arguing, theoretically, that (something like) free will is impossible; such a critique could make the empirically founded argument that humans *actually do not have anything like free will*. This kind of empirical critique would have, as far as I can tell, the same potential consequences as the theoretical critique of the framework of free will is thought to entail. The natural next question is therefore: Does such an empirically founded critique of the framework of free will exist?

If you look to the popular science literature again, you might very well get the impression that neuroscience (and other sciences studying the human body and mind) really are deciding, once and for all, this time with evidence, the age-old question of whether we have free will or not. And you would think that the answer was "no, we do not". Books like that of Harris base their argument on evidence from several fields of study, among them neuroscience. Most commonly, the neuroscientific critiques of free will have focused on the role consciousness can or cannot be said to play in action initiation and control, a tendency best exemplified by Libet and his scientific successors such as Haynes, and in the attempt at a broadly based critique of "conscious will" as seen in Daniel Wegner's *Illusion of Conscious Will* (2002). This tendency to focus on what might be termed the *causal efficacy of consciousness* (with regards to behaviour) can be understood in relation to the importance ascribed to consciousness for agency; both implicitly in commonsense thought, and explicitly in the elucidations of our commitments in legal texts and practices. Accepting that consciousness matters to our ideas of freedom and ascriptions of responsibility, investigating whether consciousness can play this role is a natural question for empirical science, not least neuroscience.

Both in Libet's early experiments (1983) and in the later work by Haynes *et al.* (Soon et al. 2008b, 2011a), showing temporal priority of subconscious over conscious processes is taken as evidence for the claim that conscious decisions come too late to be the "real" causal power driving behaviour, which therefore is ascribed to the non-conscious neural processes that are seen to predate conscious

[32] There is a theoretical possibility that neuroscience could one day evolve into a science that might have something to say about this fundamental debate, e.g. because it could somehow give (conclusive) evidence for agent causation, although I do not profess to know what such evidence would look like.

choice. The mainstay of empirically motivated critiques of "free will as framework" is therefore to deny the causal efficacy of consciousness with regard to behaviour.

However, as we shall see in Part 2 of the thesis, there are issues both with the evidence denying conscious causal efficacy, and with the way consciousness is framed in the experiments which have generated the evidence. Taking Wegner's broadly based, "revisionist" argument about the illusory nature of our experience of conscious will as a prime contender for a successful empirically founded critique of free will, I argue that the evidence he presents only supports a limited claim about part of our experience, and not a wholesale denial of causal efficacy of consciousness as this is seen to be required by our traditional understanding of free will and (moral) responsibility.

To the extent that Wegner's can be identified as the "best bet" for an empirically founded wholesale critique of free will, showing the limitations of this critique also amounts to a serious objection to this kind of approach in general. Keeping in mind the types of claims made e.g. by Harris and Haynes, this is a significant result in itself; free will is neither theoretically nor empirically refuted.

That we shouldn't expect a wholesale denial of free will and moral responsibility to arise from scientific quarters is however no reason to disregard the findings, i.a. in neuroscience, about the way humans in certain situations go about doing things in ways which appear to conflict with at least the stronger interpretations of what it means to have free will, especially as this relates to our thoughts about moral and legal responsibility. Leaving aside the question of whether free will is possible and whether any humans ever have something like it, scientific studies of the processes that lead to action and of the actions themselves can illuminate aspects of human nature that might otherwise be disregarded for reasons of theory or ideology. In other words, neuroscience and the other sciences studying humans might give critiques of the appropriateness of ascribing free will and (moral) responsibility to specific people or groups of people, in specific situations or more generally, *within* the assumed or accepted framework of a notion of free will and associated ideas of moral and legal responsibility. The limited claim I identify in Wegner's argument is significant in this respect, but for this limited claim to be incorporated into a "traditional" interpretation of the available evidence – an interpretation that is realist both about free will and about conscious causal efficacy – it is also necessary to consult the empirical evidence to see if there is support for such an interpretation; to see if consciousness really is causally effective in bringing about behaviour in any way comparable to what can be said to be required by our traditional notions of freedom and responsibility.

Part 2

Free will and the causal efficacy of consciousness

7 Neuroscience on Consciousness in the Critique of Free Will

> As humans, we like to think that our decisions are under our conscious control — that we have free will. Philosophers have debated that concept for centuries, and now Haynes and other experimental neuroscientists are raising a new challenge. They argue that consciousness of a decision may be a mere biochemical afterthought, with no influence whatsoever on a person's actions. According to this logic, they say, free will is an illusion. (Smith 2011)

7.1 Dissolving the Dilemma of Free Will

As we saw in Part I, the debate concerning free will and determinism arises out what appear to be irreconcilable facts, a "fact of experience" (*FE*) and a "fact of theory" (*FT*):

FE We experience having free will of a certain kind[33]

FT Free will of this kind appears to be impossible in (meta-) physical terms

Assuming a naïve position, there are strong intuitions to support each apparent fact: we are not usually systematically wrong about what we experience, but here we have run up against theoretical arguments that flatly contradict the particular experience we have of being free. Any thinker seeking an answer to the question "do we have free will?" is therefore faced with a dilemma: which of the intuitions should be trusted? Let us call this "the Dilemma of Free Will", or just "the Dilemma" for short. Now, I take the discussion in Part 1 to have shown *FT* to be less than certain; indeed that there is no simple reply to the question of whether free will is at all possible in (meta-) physical terms. I was motivated to embark on that analysis by the observation that some of those who offer scientifically supported arguments against free will assume that *FT* is undeniably true and therefore set their sights on undermining *FE*. Writers like Daniel Wegner thus amass evidence, i.a. from experiments like those of Libet *et al.*, to support the claim that our experience of free will is false, illusory. While Wegner pursues several lines of evidence, one aspect that we have not yet discussed is common to his entire argument: Wegner (and his fellow revisionist) argue that *consciousness* does not play the role we think it plays with regard to action, i.e. does not play the role that supposedly it would have to play if our experience of having free will of a certain kind were veridical.

This appeal to consciousness might appear entirely natural or somewhat strange, depending on your familiarity with this kind of debate. I will not enter into a long discussion of history of the idea of consciousness or the place of consciousness in commonsense and legal discourse today, but simply

[33] "A certain kind" can here be taken to refer to any of the ways in which free will has been construed in the traditional or classical debate concerning free will and determinism, what in Part I was centred around "the ability to do otherwise".

note that consciousness is strongly associated with every aspect of what it means to be an active, acting human, especially in relation to questions of whether some action was chosen or intended, and therefore with the question of when and how we are responsible, morally and legally, for the things we do.[34]

With this in mind, if we approach free will as an empirical matter, *FE* can perhaps be taken to stand for something like the following:

*FE** We have the experience of consciously initiating and controlling action.

If we assume for the moment that the experience we have is veridical, i.e. that we experience consciously initiating and controlling action because it is true that we consciously initiate and control action, *FE** could be taken to amount to a claim about what might be called the "causal efficacy of consciousness" (*CEC*) with regard to action:

CEC Consciousness is causally efficacious in initiating and controlling action.

Given the this reading of the dilemma, the road is open for revisionist arguments based in empirical research on the actual role of consciousness in action initiation, guidance and control to criticise *FE* through *CEC*. Thus, when Wegner attempts to dissolve the Dilemma by claiming that our experience of conscious will (*FE**) is an illusion, he does so based on an empirically founded argument against *CEC*.

Now, while Wegner and other revisionists set their sights on *FE* because they hold *FT* to be undeniably true, the discussion in Part 1 has excluded such simplistic appeal to *FT*. Any empirical argument against *FE* by way of *FE*/CEC* must therefore be made without relying on *FT*. To the extent that, e.g. Wegner's interpretation of evidence against *CEC* currently relies on *FT*, its potential support to an argument against *FE* will be weakened accordingly.

Another thing is now also clear: even though I have undermined the arguments to *FT* and thus any easy recourse to this in the interpretation of evidence against *CEC*/*FE**/*FE*, sufficiently strong evidence against *CEC* could still undermine our belief in free will, regardless of whether or not free will is possible in principle: If we accept that conscious control over our decisions or actions is essential for free will, evidence against conscious control would undermine free will *in practice*, i.e.

[34] I will also not attempt to give an account of what consciousness *is*: there are several interesting theories about the nature of consciousness, but none of the arguments I will review (or make) here depend on a specific theory of consciousness.

be evidence against humans actually having anything like the "ability to do otherwise" discussed in Part 1.

Finally, and correspondingly, even if the evidence against *CEC* is insufficient, this still does not give us more than a *prima facie* reason for continuing to believe in free will. Therefore, just as I provided independent reason for a continued belief in free will at the end of Part 1, I will argue here that the evidence supports a continued belief in *CEC*, which together with the argument in Part 1 can be taken as support for the veracity of *FE*.

Given this backdrop of the importance of consciousness for the empirical study of free will, we can return to the experiments by Libet and others to see what, if anything, these allow us to conclude about *CEC*/*FE**.

7.2 Conscious Causal Efficacy in Libet-Style Experiments

In the Libet *et al.* experiments discussed in Part I, subjects, based on introspection, evaluated the time at which they first became aware – consciously aware – of the urge, wish or spontaneous decision to move their finger or wrist (W). This was then compared to two objective measures, namely the onset of the characteristic "readiness potential" (RP, specifically RP II) and the onset of movement (M). When the experiments repeatedly showed the onset of RP to precede W, it was concluded that the conscious awareness of the urge or decision to move – although purportedly experienced as something like a conscious will or choice, and although reported to appear before the actual movement – could not be the *cause* of the subsequent movement. That honour was given the RP itself based on a conjunction of the measured temporal sequence and the commonly accepted idea that a cause must precede its effect. Since RP appeared consistently before both W and M, the conclusion was drawn that RP must be the real cause of M (RP $\rightarrow$ M), with W either a product of RP or something produced in parallel with the causally effective processes which lead from RP to M (RP $\rightarrow$ [W$\rightarrow$] M).[35] Thus, Libet's results were taken by many to be unequivocal proof of the claim that consciousness does not initiate action, thereby denying *CEC*.

Libet and his colleagues first suggested and later argued that there was a gap of around 100 ms in which one could *consciously veto* the movement about to be made (Libet 1998, Libet et al. 1983). According to this picture, conscious will would not be the initiating cause of action, but would be able to serve as a final "controller" of action execution, which in normal situations would arguably

[35] By way of illustration, the various interpretive possibilities identified by (Sinnott-Armstrong 2011) are shown here in a simplified annotation with brackets surrounding W, indicating the uncertain relation between W and M, but with RP given temporal and causal priority in every case.

translate to the kind of ultimate "executive" control needed for ascriptions of responsibility. However, most critics dismissed this suggestion as a non-starter, and reasoned that parity with the (RP → [W→] M) argument would imply that the conscious "veto" itself was also caused by preceding brain activity, relegating any veto power back to those unconscious mental mechanisms.

The experiments by Haynes *et al.* (Soon et al. 2008b, Haynes 2011a) have been presented as evidence to strengthen the claim that the real causal processes leading to decisions and behaviour are physical and non-conscious (Smith 2011). In Haynes' Libet-style experiments, subsequent choice between pressing a button connected to the left finger or one connected to the right was predictable with up to 60% accuracy from data recorded a full 6-7 seconds before the subject actually pressed one of the buttons. Taking into account the time lag inherent in the method used to measure activity (BOLD fMRI)[36], this has been taken to mean predictive patterns of activity in areas of the cortex up to 10 s before conscious awareness of choice. Appearing to show non-conscious neural processes "deciding", not only the 'when', but also the 'what' of action at a subjectively appreciable interval before the person choosing has any experience of actually making the decision, these results offer intuitively compelling evidence against what might be called "conscious free will". Libet-style experiments have thus reanimated the above quoted belief that consciousness or conscious will is a mere "biological afterthought", an *epiphenomenon*[37] hitching a ride on the inexorable motion of the physical nervous system.

As for consequences for the debate concerning free will, if *FE*/CEC* is accepted as the correct interpretation of the entailments of *FE*, or alternately, simply that consciousness in the role in which it is portrayed by *FE**, i.e. *CEC*, is important to free will and moral and legal responsibility; and if we agree that Libet's experiments and subsequent studies modelled on these show that consciousness comes too late to be the cause of action, then it would seem that the dilemma is dissolved, and furthermore, that we have a successful wholesale scientific denial of free will.

---

[36] Blood Oxygen Level Dependent functional Magnetic Resonance Imaging is a brain-imaging technique that relies on the difference in magnetic field properties of haemoglobin carrying oxygen compared to haemoglobin without oxygen to indicate the changes in neuronal activity in areas of the brain (reasoning that neurons that are more active require a larger local blood flow to supply nutrients, but have little or no increase in oxygen consumption, thus giving a net increase in blood oxygen level in the area of activation). This technique has high spatial resolution (2-3 mm), but poor temporal resolution (around two measurements per second, and lagging several seconds behind neural activity) compared to, e.g. EEG (on the order of centimetres and milliseconds), and is heavily reliant on statistical analysis in order to get useful results. This also makes results from fMRI controversial, since it is possible to get widely varying interpretations based, i.a. on how one generates the "baseline" activity against which task-relevant activity is measured (Stark and Squire 2001, Gusnard and Raichle 2001). For more reasons to be generally sceptical of neuroscientific results, see (Button et al. 2013)

[37] A phenomenon with no causal effects; caused but not causing. The idea or worry that consciousness is an epiphenomenon has important historical roots, as we shall see in section 9.2

However, there are issues with both premises:

1. Construing conscious control in accord with *FE*/CEC*, as exemplified by Libet-style experiments including those by Haynes *et al.*, is not the only way of accounting for the role of consciousness in the generation of behaviour, nor is it, on closer inspection, necessarily even a good one. While it might seem intuitively right to identify any conscious causal power with the direct initiation and control of action, this actually accords rather poorly with the role consciousness is commonly taken to play in complex behaviour of the kind where questions of freedom and responsibility are commonly raised, where goal-directed behaviour comprising a wider temporal and thematic range of thoughts and intentions are essential to understanding action (Gallagher 2006).
2. Even if one were to accept this construal of the role of consciousness, there remains two serious issues with the inference from the kind of experiment performed here to a general truth about choice and action:
    a. There are good reasons to be sceptical both of the claim that W is anything like a conscious decision or act of will, and of the assumption that RP is the neural cause of action, as showed by experiments where change in instructions or merely in analysis produces results which are inconsistent with the claims of Libet *et al.* (Pockett and Purdy 2011).
    b. Even if we were to accept that W really corresponds to what we think "will" is and that RP is the real cause of movement in Libet's studies, this is still only valid for the kind of highly constrained and highly artificial experimental setting used in Libet-style experiments (Bayne 2011).

Whereas (1.) is mainly a conceptual worry related to the framing of the empirically informed debate concerning free will and conscious causal efficacy, (2.) concerns the evidence itself as it is presented within the given conceptual framework. In order to criticise fairly the extant empirical evidence, I will allow the assumed framework for now, and return to the possibility of a different framing of consciousness in chapter 9.

Issue (2.a) is a serious objection to the claims made by Libet *et al.* as well as any subsequent experiment that similarly relies either on measurements of RP, a subjectively evaluated time of (spontaneous) "decision" (W) based on the instruction to wait for an "urge" to act, or both. While it is possible (and common) to object to the reliance on the ability of experimental subjects to accurately evaluate the time of W (Pockett and Purdy 2011, pp. 29-30), there is a more

fundamental objection to be made about the claim that W is anything like a conscious decision to move. In most Libet-style experiments, including those by Haynes *et al.*, subjects are instructed to wait for an "urge" to move. Ostensibly worded thus to avoid pre-planning of action and transference of effects from one trial to the next, guaranteeing that each measured "choice" really is "spontaneous" (Libet et al. 1983, Soon et al. 2008b), these instructions also mean that subjects are not focused on making a decision of when to move or what button to push, but rather direct their attention towards detecting the arrival (from somewhere) of an urge to move (Batthyány 2009). The problem with this is two-fold. Firstly (2.a.a), subjects are arguably not engaged in making decisions or choices at all, and thus the relevance of the experiments for free will, understood as freedom in decision/choice, is undermined. Secondly (2.a.b), even if the wording is supposed to pre-empt pre-planning of action, there is an argument to be made that subjects in Libet-style experiments rely on (implicit) plans for *how* they should act during the experiment, e.g. how often they should press the button so as to be good experimental subjects (Talmi and Frith 2011), and that this makes difficult the conclusion that measured brain activity really precedes the conscious choice to act.

Considering (2.a.a), one of the most significant critiques of Libet's original experiments has been provided by Susan Pockett and Suzanne Purdy (Pockett and Purdy 2011). Using a combination of reanalysis of extant experimental data and evidence from new experiments, Pockett and Purdy were able to show i) both that Libet's type II RPs are neither necessary nor sufficient for spontaneous voluntary movement, and that RPs are probably a subclass of a signal related to "general readiness" or expectancy rather than the preparation of movement. This is further supported ii) by the fact that new experiments replicate Libet's only when subjects are instructed to wait for an urge to move; for trials where subjects are instructed instead to make an explicit decision to move, either no RP or only a much shorter RP that coincides with or arises after the time of decision is detectable. This clearly undermines the argument that classic Libet-style experiments show conscious choice coming too late to be the real cause or initiator of movement.

Haynes *et al.* are mindful of the possibility that increased activity in certain brain areas leading up to a decision or choice may only be "unspecific preparatory activation" (Soon et al. 2008b, p. 543), reflecting, e.g. "attention to intention" (as Hakwan Lau et al. put it, Lau et al. 2004), rather than a pre-conscious decision to move; but argue that the measurement of activity predictive of one of two possible choices – as in their experiments – circumvents this objection. Thus, showing that patterns of activation of an area of the brain predicts with 60% certainty which of two buttons will subsequently be pressed, up to 10 seconds before the time at which the subject is aware of making a choice between the buttons (Haynes 2011b, Haynes 2011a), could perhaps be taken as evidence of

the claim that consciousness "comes too late" to be that which decides what we do. The choice is already encoded in certain brain areas long before we become aware of it.

There are, however, significant hurdles to such a conclusion. On the one hand, the evidence obtained does not support it. While it is certainly an impressive feat to decode patterns of activation that can predict subsequent choice with 60% certainty, this is still a far cry from the claim that your brain decides all your choices 10 seconds in advance of you yourself becoming aware of them – a point Haynes himself duly notes:

> Importantly, a different interpretation could be that the inaccuracy simply reflects the fact that the early neural processes might *in principle* simply not be fully, but only partially predictive of the outcome of the decision. In this view, even the full knowledge of the state of activity of populations of neurons in frontopolar cortex and in the precuneus would not permit to fully predict the decision. In that case the signals have the form of a biasing signal that influences the decision to a degree, but additional influences at later time point might still play a role in shaping the decision. Until a perfect predictive accuracy has been reached in an experiment, both interpretations – incomplete prediction and incomplete determination – remain possible. (Haynes 2011b, p. 93)

As is often the case, Hayne's is not as bombastic in his peer-reviewed articles as his popularisers are when making claims from the experimental evidence (Harris). Nevertheless, it is probably not uncharitable to say that Haynes himself is leaning towards the "revisionist" interpretation that our conscious choices are, if not fully, then at least importantly determined (in advance) by non-conscious brain processes. Whether made by Haynes himself or by his popularisers, the revisionist interpretation must be tempered when considering the limitations inherent in the Libet-style experimental design that is still employed, albeit in modified form, by Haynes *et al*. Thus, while they rely on a series of letters on a screen rather than an "oscilloscope clock" to allow subjects more accurately to time their conscious choice, Haynes *et al*. still ask subjects to wait for an urge to move to spontaneously arise, which – although they argue that prediction of choice between options circumvents the objection that the measured activity is merely generally preparatory or activation-in-expectation (of something) – still leaves open the objection that subjects *are* skewed towards waiting and "introspecting" for a something to appear, rather than making any real choice or decision. Thus, whatever the predictive power achieved by Haynes, his subjects are arguably not deciding when these predictive patterns emerge, but are simply waiting for an urge to push a specific button to arise on its own. This objection can be expanded and strengthened if it is connected to the second part-objection identified above (2.a.b), namely that subjects in Libet-style experiments – although given instructions which are meant to pre-empt preplanning, etc. – might

nevertheless be shaped in their behaviour by more or less explicit ideas about the expectations of others (most notably those of the experimenter), or of their own expectations about being subjects in an experiment, e.g. the thought that being a good experimental subject entails following instructions; and that this, for Libet-style experiments, actually requires you to feel an urge, or at the very least, to push the button at least once, and probably a number of times, as well as making sure that the button-presses are spaced out appropriately (Talmi and Frith point out this often ignored aspect of the experimental situation, although they focus on the issues associated with the implicit suggestion that experimental subjects should feel an urge to move, Talmi and Frith 2011).[38] Any subject who fails to comply with these more or less explicit requirements is likely to be corrected by the experimenters (otherwise, the data will be useless), and so it seems not entirely speculative to say that subjects are probably shaped in their behaviour by more or less explicit attention to these requirements, even if they never have to be (explicitly) corrected. This arguably also touches the instruction, given to subjects in Haynes' experiments, to choose randomly *which* button to push – although Haynes takes pains to provide observations indicative of the true randomness and autonomy of his subjects' choices, with a statistical analysis of button sequence showing a nicely random distribution (Soon et al. 2008a) despite not instructing his subjects to distribute their responses (Haynes 2011b, pp. 91-2). However, as he points out in another article:

> Importantly, in order to facilitate spontaneous behavior, we did not ask subjects to balance the left and right button selections in successive trials. This would require keeping track of the distribution of button selections in memory and would also encourage preplanning of choices. *Instead, we selected subjects that spontaneously chose a balanced number of left and right button presses without prior instruction based on a behavioral selection test before scanning.* (Haynes 2011a, p. 12, emphasis mine)

Thus, there is a live possibility that Haynes' subjects, specifically selected for the trials that provide his evidence because they "spontaneously chose a balanced number of left and right button presses", are simply (implicitly) aware of the niceness of having an even distribution of responses, and act accordingly. This could undermine the validity of the evidence, since the combination of waiting for an urge to act and the (implicit) ideas that can or do shape both these urges and the actions that are to result from them, means that any patterns of brain activation might simply reflect the various

[38] Libet et al. were aware of this possibility: "Asking subjects about 'surprise' acts should have indicated to them that it was acceptable even to have and to report the absence of a conscious urge or intention to act prior to a self-initiated act. The fact that instances of 'surprises' were reported increases confidence that the reports of timing prior to the act represented endogenous experiences not defined or induced by the instructions." (Libet et al. 1983, p. 627) However, this is only a guess on their part, and, in any case, does not address the worry arising from the implicit requirement to perform (a number of) self-initiated button presses.

biases that shape subjects' choices over time, e.g. the bias towards an even distribution of left and right button presses shaping future choice in light of past button presses. If nothing else, this is potentially a significant source of error.

Finally, as noted in (2.b), even if the assumptions made by Libet and others about the role of measured preceding brain activity and the casting of 'W' as a conscious choice to move are granted, the undeniable simplicity of the experimental task represents another hurdle for those who wish to make a general claim about the causal efficacy of consciousness from these experiments. Lying still and pressing a button is not a good paradigm even for simple tasks in real life; even less so if the tasks with which we are concerned are actions with potential immediate moral implications (Roskies 2006b, Roskies 2012, Roskies 2011).

In sum, there is a long way to go from the kind of evidence presented by Libet and his successors to the claim that consciousness is causally inefficacious in bringing about behaviour, and thus *FE*/CEC* still stand as plausible.

That the validity of evidence from Libet-style experiments have been so extensively questioned is probably one of the reasons why psychologist Daniel Wegner's comprehensive treatment of the question of free will in his 2002 book *The Illusion of Conscious Will* (hereafter ICW, Wegner 2002) has gained such influence with the "revisionist" side of the debate. Wegner is now often mentioned alongside Libet as one of the most influential thinkers to offer scientific support for the argument that consciousness cannot and does not play the role we normally attribute to it in action initiation and control, i.e. that *FE** and *CEC* are false (Even Pockett and Purdy conclude their article by expressing belief in Wegner line of argument (Pockett and Purdy 2011). Building partially on Libet and others, but adducing a much wider range of experiments alongside varied evidence from both pathological and non-pathologic examples of dissociations between what he terms *phenomenal* and *empirical* will, Wegner argues that what we usually take at face value – our experience of consciously causing, initiating or controlling our actions ("phenomenal will") – is an illusion, i.e. does not reflect the actual mechanisms that bring about action ("empirical will"). In other words, he attacks *FE** and attempts to dissolve the dilemma by denying that our experience of consciously initiating and controlling action is *veridical* – hence illusory. Wegner's version of the illusion argument is widely based and well developed, and is probably one of the best bets for a wholesale critique of the framework of free will to date. But is it successful?

8 Is Conscious Will an Illusion (with a Purpose)?

> "Conscious will is the mind's compass. […] The experience of will is therefore an indicator, one of those gauges on the control panel to which we refer as we steer. Like a compass reading, the feeling of doing tells us something about the operation of the ship. But also like a compass reading, this information must be understood as a conscious experience, a candidate for the dreaded "epiphenomenon" label. Just as compass readings do not steer the boat, conscious experiences of will do not cause human actions." Daniel Wegner (2002)

8.1 Wegner's Argument

Libet and others using Libet-style laboratory experiments have not been able to provide convincing evidence to refute *CEC*/*FE**. The data is too weak, the assumptions made in acquiring it are unsupported, and the experiments themselves have serious methodological problems. They might be on to something, but they are not there yet in terms of evidence or explanatory power. In addition to this, there is the question of why we are equipped with an experience of free will if there is nothing to it. While classical epiphenomenalists do not seem overly bothered by the sheer uselessness of the phenomena to which they assign this label, it is quite reasonable for normal people to wonder why evolution, notoriously fixed on function, would equip humans with the illusory experience of being in (conscious) control if it did not serve any purpose.

Wegner's book *The Illusion of Conscious Will* (Wegner 2002) can be seen as an attempt both at remedying the weakness of the evidence from Libet-style experiments, and at giving a reply to the nagging question of what an illusion of free will would be good for. The premise of Wegner's book is that there is evidence to the effect that we can be mistaken about our conscious control over something, both attributing control to ourselves when in fact the action in question was performed by someone or something else, and failing to recognise that we are responsible for something we have in fact done or contributed to. There is, that is, evidence both of "illusions of control" and of "automatisms". From this apparent dissociation of phenomena, Wegner argues for what he terms the "illusion of conscious will", i.e. a substantial dissociation between the conscious experience of will and the actual mental mechanisms that bring about action. As for why humans are subject to such an illusion, Wegner argues that what might more properly be called the "feeling of authorship" is an important element of social coordination and character formation. Thus, while the illusion of conscious (free) will is an epiphenomenon in terms action initiation and control, therewith denying *CEC*; *FE* or *FE** are not completely epiphenomenal, and therefore not as hard to explain (away) as in traditional epiphenomenalism.

8.1.1 Automatisms: Unexpected Absence of Conscious Will

As Wegner uses the term, an "automatism" is any "apparently voluntary behaviour that one does not sense as voluntary."(Wegner 2002, p. 9, footnote 3).[39] Evidence for automatisms comes from varied sources. There are the strange cases of "alien hand syndrome", in which a person will attest to the fact that one of her or his arms is acting "on its own"; apparently purposefully ("quite willful" as Wegner puts it, 2002, p. 5), but without any sense of voluntariness or agency, and often at cross purposes to the person who's hand it is, sabotaging the actions s/he is trying to perform with the other hand. Considering hypnosis, Wegner argues that the most salient quality of behaviour resulting from (post-)hypnotic suggestion is the lack of wilfulness that accompanies it; behaviour that by all other measures is attributable to the person who has been hypnotised, but which the person acting rather feels is happening to her (Wegner 2002, pp. 6, 271ff). Wegner also brings to the fore historic examples of automatisms, including the *table-turning* fad that arose in spiritualist milieu in England and America in the mid-nineteenth century: In a session of table turning, a group of people would sit around a round table with their hands on the table top, ostensibly doing nothing, when the table would suddenly start to move, often in a "sinister" fashion, typically indicating the presence of hidden dark powers (Wegner 2002, pp. 7-9, 100ff). Of course, and as proved at the time by the famous physicist Faraday, the actual cause of the table's movement was the people sitting around it. Another historic example with spiritualist credentials is automatic writing: a person, more or less practiced in the feat, sits down at a writing table, hand resting on a small plate, which in turn rests on the table supported by two wheels and a pencil tip; and scribbles away. After a while, strange words, meaningful sentences or even whole paragraphs can appear, without the person physically writing reporting any sense of being the author of those phrases (Wegner 2002, pp. 103ff).

In a less frivolous vein we find the controversy surrounding so-called "facilitated communication" (Wegner 2002, pp. 195ff): based on the irreproachably good intention to facilitate the communication of persons with poor or no powers of communication (initially children with cerebral palsy, later also severely autistic children), trained facilitators (typically) support (or rest on top of) the arm or hand of the person they are facilitating, helping them to type on a keyboard. Initially, the technique was deemed a great success when children who formerly had been non-communicative could suddenly express themselves in writing quite comprehensibly and intelligently, in one famous case moving on later to successfully argue in a courtroom for her release from hospital (Crossley and McDonald 1984). The technique was however controversial, and became all

[39] Note that this is markedly different from how 'automatism' is used in legal contexts, as discussed above.

the more so when several peer-reviewed experiments repeatedly showed alleged "communication" rather to be the direct product of the facilitator (Jacobson, Mulick, and Schwartz 1995, Mostert 2001). That's not to say that the facilitators knowingly took over or led those they were trying to help, quite the contrary: facilitators had no experience of their facilitation slipping into control, even if this was what was really happening in most of the cases. While anecdotal evidence supports the claim that at least some forms or instances of facilitated communication constitute genuine communication from the person facilitated, the demonstrably false cases provide, at the very least, yet another dissociation between being the physical cause or author – here of something as complex as written communication –, and the feeling or experience-based conviction that one really is responsible, which we would normally expect to accompany such action.

8.1.2 Illusions of Control: Unfounded Experience of Conscious Will

As for "illusions of control", examples are neither as numerous nor striking, but there has been some work on the phenomena in social psychology (see e.g. Thompson, Armstrong, and Thomas 1998), and Wegner argues that we quite commonly experience varying degrees of this illusion e.g. when we operate the buttons of an elevator or a vending machine and we are unsure whether our action has caused anything to happen. He also argues that we experience an illusion of control in cases where we (superstitiously) think that our actions affect our surroundings in physically impossible ways, e.g. "jinxing" a sporting event on TV by something trivial like leaving the room (Wegner 2002, p. 10).

Humans are indeed susceptible to superstitious belief in mysterious paths of influence, as are other animals who can learn by conditioning. B.F. Skinner, the famous "behaviourist" psychologist who suggested we do away with mental explanations altogether, showed this with pigeons (Morse and Skinner 1957, Skinner 1948): after having tested regular conditioning, in which the pigeon would perform a certain act and reliably get a treat, Skinner changed the feeding schedule to random intervals independent of the pigeons' actions. As a result, the pigeons would repeat whatever random behaviour they had been performing at some point when they had been fed, e.g. moving around in a circle.

Although the results are contested (and behaviourism mostly abandoned), they do bear a certain resemblance (that has also been investigated empirically, see Rudski 2001) to the human propensity to stick with what has "worked" in the past, even if "sticking with what works" entails doing things where no plausible physical account can be given for their effect. This is especially prominent in high stakes settings such as professional sports (or in war), situations where " […] chance can sometimes defy even the most rigorous training and discipline", as author Elizabeth D. Samet says

of baseball, after becoming a "lucky charm" for her West Point academy team (Samet 2010). Thus, sportswomen and -men might have "lucky" underwear, a convoluted pregame ritual or other things they have to do in order to maximise their (perceived) chances of success. Of course, for humans, once such a ritual or talisman has been adopted, it does in fact become directly influential in the sense that its disruption might disrupt the psychological part of the performance that follows. Still, the actual path of influence here is arguably different from whatever "lucky" power originally attributed to the thing in question, and so one might still say that this constitutes an illusion of control.[40]

8.1.3 Significance for Free Will Debate

If we allow that these examples provide tentative evidence for a *double dissociation* between the experience of "consciously willing" and actually doing something, two responses seem immediately available: either we could focus on the fact that these situations all appear to be exceptions to what we commonly (and supposedly correctly) experience as a robust connection between our choices and our actions. Or, says Wegner, we can focus on just how easily we are misled about something so basic as our own agency: If it really were the case that our experience of conscious will was directly connected to (or identical with) the processes that generate action, we would arguably not expect such blatant errors as evinced in these examples. Asking us to allow for the possibility that our ordinary perceptions may be mistaken, Wegner presents the hypothesis that our experience of conscious will is not a reflection or direct perception of actual agency at all, but rather something generated alongside the actual causal non-conscious mental processes that issue in behaviour. If this is the correct understanding our experience of consciously "willing" things, this experience would arguably be illusory since it is not a veridical reflection of the real causal path developed subconsciously.

The significance of this for the free will debate is clear in Wegner's project. He introduces his book with a version of the Dilemma, in which the opposing explanatory modes of a mechanistic science and a personal belief in conscious will are seen to have "an oil and water relationship" (Wegner 2002, p. 2). His proposal is that we can resolve this apparent conflict if we can explain how and why we have the experience of conscious (apparently free) will despite the fact that free will in the traditional sense is incompatible with a scientific understanding of human action:

[40] Or perhaps more appropriately a *delusion* of control – although it is far from clear that people might not be aware of both sides of this while still continuing with the practice, so that there really isn't an delusion at all, but rather a voluntary endorsed belief in something one knows does not quite fit in with other commitments.

> One way to put [the two] together—the way this book explores—is to say that the mechanistic approach is the explanation preferred for scientific purposes but that the person's experience of conscious will is utterly convincing and important to the person *and so must be understood scientifically as well*. (Wegner 2002, p. 2, emphasis mine)

Showing that the experience of will is sometimes dissociated from actual causal processes that issue in action is the first step towards substantiating the claim that it is always so, and this in turn is the basis for Wegner's scientifically based resolution of the Dilemma, by way of denial of *CEC*/*FE**.

8.1.4 Apparent Mental Causation

Central to Wegner's argument in *ICW* is the idea that our experience of consciously willing or controlling something is constructed by subconscious mental mechanisms based on the fulfilment of a set of criteria. Comparable to the way "we" automatically and subconsciously infer causality between external objects when their relative movements fulfil certain criteria, we automatically and subconsciously infer or interpret that our thoughts are the causes of certain happenings (typically actions) when the two coincide in a particular fashion (Wegner 2002, pp. 64ff). For there to be an experience of conscious will, i) our thoughts about an event must occur prior to the event, ii) the same kind of though must consistently occur with the same kind of event, and iii) there must be a relative absence of other factors to explain the event. These are the priority, consistency and exclusivity criteria that Wegner argues must be met for any experience of will to arise. When thoughts appear thus reliably as "previews" of upcoming action, we have the experience of consciously willing those actions.

In an analogy with stage magic, Wegner says that our experience of conscious will – "phenomenal will" – arises out of the *perceived causal sequence* of thoughts leading to action. As with stage magic, the *real causal sequence* is hidden from view, far more complex (and not at all magical); and will not usually or necessarily correspond to the one that is perceived. In the cases where "we" – our thoughts or mental states, that is – actually *are* the causes of action through "a massively complicated set of mechanisms", this is a manifestation of "empirical will" (Wegner 2002, pp. 26-7). The illusion of conscious will is then essentially the result of mistaking phenomenal will for empirical will (Levy 2007, p. 229).

Wegner calls this the "theory of apparent mental causation":

> *People experience conscious will when they interpret their own thoughts as the cause of their actions*. […] This means that people experience conscious will quite independently of any actual causal connection between their thoughts and their actions. Reductions in the impression that there is a

> link between thought and action may explain […] automatisms […]. And inflated perceptions of the link […] may, in turn, explain why people experience an illusion of conscious will at all. (Wegner 2002, p. 64, emphasis in original)

In other words: with every experience of conscious will being the product exclusively of interpretation, abnormal cases of too much or too little "will" can be explained as easily as the standard, appropriate-interpretation case(s), namely by case by case analysis of the basis for the interpretation.

8.1.5 The Emotion of Authorship

Wegner makes a convincing argument that the feeling we may or may not have of having caused something to happen is neither the same thing as whatever mechanisms actually bring about action, nor an infallible readout of the workings of these mechanisms. But while our experience of consciously causing or being in control over our actions is illusory, the experience itself is not really an epiphenomenon. Rather, argues Wegner, our experience of conscious will serves a different function in our lives as acting subjects, namely as a sort of "authorship emotion" (2002, pp. 325ff) that may vary in strength to support anything from perfect certainty that something was done by ourselves to perfect certainty of the contrary. While this authorship emotion most often arises in connection with acts of which we really are the author, it can also fail to appear in appropriate circumstances, or also arise in inappropriate ones. Since the "emotion of authorship" often *is* correctly attached to actions performed by ourselves, it serves as a "body-based signature" of the actions that are likely to be our own, allowing us "to develop a sense of who we are and are not", "to set aside our achievements from the things that we cannot do", and "[allow] us to maintain the sense of responsibility for our actions that serves as a basis for morality." (2002, pp. 325, 327, 328). The nagging question about the point of an illusion of free will is therefore answered on an altogether positive note:

> The fact is, it seems to each of us that we have conscious will. It seems we have selves. It seems we have minds. It seems we are agents. It seems we cause what we do. Although it is sobering and ultimately accurate to call all this an illusion, it is a mistake to conclude that the illusory is trivial. On the contrary, the illusions piled atop apparent mental causation are the building blocks of human psychology and social life. It is only with the feeling of conscious will that we can begin to solve the problems of knowing who we are as individuals, of discerning what we can and cannot do, and of judging ourselves morally right or wrong for what we have done. (2002, pp. 341-2)

8.2 Critiquing the Argument: Interpreting the Experimental Evidence

In addition to what has already been mentioned about automatisms and illusions of control, Wegner brings to the fore a wealth of empirical and anecdotal evidence to support the idea that "conscious will" is illusory, or rather, to substantiate the theory of apparent mental causation. The argument runs an integrated course throughout the book, drawing upon many strands of research and more or less closely associated ideas, such as the Chevereul Pendulum and dowsing; ideomotor action and ironic effects (phenomena where conscious or subconscious influences related to action unintentionally causes associated movements); the "ideal agent" hypothesis (that we have an idea of an ideal agent who always acts for appropriate reasons, and that we are therefore liable to invent reasons or intentions after the fact if we find ourselves having); the theory of the "left hemisphere interpreter" (who's task it is to make sense of things, i.a. what the right hemisphere is doing, and who's interpretations can become pure confabulation when dissociated from the processes that cause action); the existence of action projection (where agency is projected to outside forces, e.g. other people or invented, often supernatural agents to explain phenomena); invasive thoughts in schizophrenia (where i.a. speech that is probably self-generated is mistaken for taunts, threats and commands from one or several external agents); and so on.

Wegner discusses far too many interesting phenomena and too much experimental evidence to be presented or commented upon here. Taken as a whole, the evidence presents a strong case for the claim that our experience of being acting agents, our attribution of agency to ourselves and others, and our understanding of the causes of our actions *can* be misleading relative to what would be determined from an objective point of view. There is, however, good reason to temper the conclusion Wegner reaches from this.

Apart from what might be said about the abnormal presence or absence of an experience of will, all of the studied situations are abnormal in one way or other, either involving pathologies (as with schizophrenia and alien hand syndrome), some degree of self-deception or complicity (as in spiritualism and hypnosis) or experimental trickery specifically intended to deceive or bring about error. Furthermore, in the controlled experimental conditions, which one would think would provide the best or most "clean" evidence for the dissociation, the experience of conscious will is only ever modulated (as in the oft-cited "I Spy" study, see below), and trickery is discovered by the subjects if the discrepancies introduced are too large (in line-drawing experiments with false feedback, subjects detect deviations between input and feedback above 14°, Pacherie 2006).

Take, as an example, the "I Spy" study that Wegner uses as evidence for the theory of apparent mental causation (Wegner 2002, pp. 74ff). The experimental setup consists of two people with headphones jointly moving a computer mouse in circles by resting their fingers on opposite sides of a plate attached to the mouse (similar to the operation of an Ouija board). The mouse pointer moves correspondingly on-screen where an image containing a myriad of small items is displayed. For each trial, after some time of moving the mouse around in this manner, music will be played over the headphones, with the participants previously having been instructed to stop the mouse sometime during this music. Unbeknownst to the real experimental subject, her fellow participant is actually an experimental confederate, whom for certain trials is instructed over the headphones where she should force the pointer to stop. In addition, for some trials, the real subject will hear a word denoting one of the figures on screen at variable intervals before, during or after the music is played.

Figure 3: "Participant and confederate move the mouse pointer together in the I Spy experiment by Wegner and Wheatly (1999)" Source: (image and caption from Wegner 2002, original in Wegner and Wheatley 1999)

The parameters are varied across trials, so that on some trials, the subject will hear a word denoting a certain object on screen (e.g. "Swan"), and the confederate will (shortly after) be instructed to stop the pointer near the swan. Compared to trials where the parameters are different ("Swan" is not said, or is said long before or after stopping the mouse; the confederate is not instructed to stop), this particular configuration tests the hypothesis, based on the theory of apparent mental causation, that eliciting thoughts which are consistent with a later event just prior to that event will augment the experience of conscious will in the subject. The experimental results support the hypothesis, with subjects reporting a level of intentionality up to about 60% on forced-stop trials where e.g. "swan" was said 1 second before a stop, compared to a mean of 52% for forced-stop trials on average and a mean of 56% for the rest of the trials, in which the confederate was instructed to let the participant decide.[41]

However, while this does support the hypothesis that thoughts consistent with the result elicited (externally) just prior to the result increases the experience of the result being intended, it does not support the hypothesis that the "experience of conscious will" is purely the result of inference, nor that it is (somehow) illusory. Subjects in the "I Spy" study reported the result being intended (albeit to a lesser degree) also on other trials, and while the confederate forced the pointer to stop, the subject is not mistaken about acting (she is indeed moving the pointer around, and has presumably chosen to do so); nor can it be said that the subject does not contribute anything to the stop, since the confederate would (likely) be hard-pressed to force a stop if the subject was unwilling.[42, 43]

In his review of *ICW* (2002), Eddy Nahmias points out that the anecdotal and experimental evidence Wegner presents can all be seen as exceptions to the general rule of correspondence between "conscious will" and action. This is arguably also the most common interpretation, with the cases Wegner presents being notable just because they are such exceptions. The fact that there are exceptions does not show that conscious will is causally irrelevant in the standard case of correspondence; it "only" shows that the "experience of conscious will" is neither necessary nor sufficient for (voluntary) action. This in itself is a significant result, but Wegner wants to go further.

---

[41] 1% corresponded with the statement "I allowed the stop to happen" and 100% with "I intended to make the stop". Note that "allowing to happen" is also not without its significance as far as behaviour (in the non-specialist sense) is concerned.

[42] Due to the experimental setup, the last of the three criteria – the criterion of exclusivity – cannot be varied, but will (according to the theory of apparent mental causation) be a constant source of attenuation of conscious experience of will, thereby (presumably) explaining the average figures across the board (the participant is unsure or confused). It would be interesting to see the results of a single-subject variation on this experiment, where a forced pointer stop was accomplished in some way that minimized the perception of external agency in the subject (perhaps accomplishable with careful, gradual application of electromagnetic force to the mouse).

[43] This cooperative aspect of the experiment arguably complicates analysis of individual action, with so-called "contributive acts" to cooperative action being the subject of a whole field of study in social ontology (see e.g.Schmid 2005).

Aware of the standard interpretation, Wegner words his arguments tentatively, but the claim is still that what we think are exceptions are actually the rule – that automatisms reveal the real, causally effective processes that underlie action as they are *sans* the illusion of conscious will:

> Automatisms could flow from the same sources as voluntary action and yet have achieved renown as oddities because each one has some special quirk that makes it difficult to imbue with the usual illusion of conscious will. Automatism and ideomotor action may be windows on true mental causation as it occurs without apparent mental causation. (Wegner 2002, p. 130)

But, as Nahmias points out:

> [A]nother interpretation is that automatisms are *not* produced by the same kinds of processes that create intentional action precisely because the causal role of conscious intention has been bypassed. (Nahmias 2002, p. 533)

There are, in other words and as already mentioned, at least two possible interpretations of the evidence presented by Wegner. The standard or traditional interpretation is that all these automatisms and illusions of control are exceptions to the rule of conscious causal efficacy – i.e. that automatisms are *different* kinds of action compared to regular, voluntary action; that ordinary conscious processes are "out of the loop" only in these special cases, and that this is what makes them special. Wegner's alternative, revisionist interpretation is that automatisms show action generation as it really is, stripped of the illusory experience of conscious causal efficacy, and correspondingly, that illusions of control show this experience devoid of its supposed correlate, the actual causal connection between agents and acts. Consciousness is therefore always "out of the loop", and we think otherwise only because we have been misled by appearances.

Given the set of evidence presented by Wegner, the traditional and the revisionist interpretation can be seen as instances of "inference to the best explanation" – competing hypotheses that can account for the same experimental data. If we accept this framing of the issue, adjudicating between the two interpretations becomes a matter of evaluating *the way* they explain the observed phenomena, i.e. their relative "explanatory virtue".[44] Wegner takes the revisionist interpretation to provide the best explanation, but the traditional interpretation is far from obviously inferior. Indeed, the revisionist explanation must arguably cover a wider explanatory stretch since it is arguing for a reversal in our

[44] Inference to the best explanation, also known as *abduction*, is at once a very common and controversial type of inference. While we humans are usually adept at evaluating which of two competing explanations is the best (e.g. in the case of whether the room is still messy because you didn't clean it as you said you would, or because a poltergeist messed it up again after you had cleaned it), giving a precise account of just what considerations should matter when we say that one explanation is "better" than the other – especially as one starts to consider difficult cases in science – is far from trivial. For more on this, see e.g. (Douven 2011).

view of normal and exceptional cases, effectively making the most common case – the experience of consciously initiating and controlling actions that are *actually* initiated and control by ourselves – the one in need of explanation. Evaluated side by side, the revisionist interpretation thus appears somewhat unmotivated given the extant option to explain the *prima facie* exceptions of automatisms and illusions of control *as* exceptions to the standard case of causal efficacy of consciousness. The traditional explanation seems unquestionably to be the "better" of the two, given the set of data presented by Wegner,[45] and in comparison, the revisionist interpretation appears more like a purely sceptical challenge, put forth only because it is possible.

This, however, misses one important point. Wegner's revisionist interpretation is not only motivated by the empirical evidence presented in the book: it is, as shown, an attempt at dissolving *the Dilemma* – a dilemma that from the point of view of Wegner is perpetuated by the traditional, "realist" interpretation of conscious causal efficacy, and which would be dissolved given the acceptance of the revisionist interpretation he presents. However, given what was shown in Part I about the difficulty of substantiating the "fact of theory" *FT* that free will appears to be impossible, such motivation cannot be said to be *objectively* present. Consequently, Wegner's revisionist interpretation must be considered a challenge to *FE** based wholly on its merits as an explanation of empirical data.

The traditional interpretation thus retains its *prima facie* advantage as the better explanation of the considered evidence. That being said, we should not be content with leaving the question open like this, since a simple appeal to common sense and extant divisions into rules and exceptions is unsatisfactory in face of the wealth of empirical evidence presented by Wegner about how surprisingly often we appear to rely on inference or interpretation to arrive at our judgements of agency and responsibility, and how we may misjudge the level of our own contribution to some observed effect. If our "experience of conscious will" were straightforwardly connected to whatever other systems that underpin agency, such rates of error would not be expected. If automatisms and illusions of control really are the exceptions, why do they arise at all? The traditional interpretation has an important *lacuna* in this respect, and so we should take the challenge presented by Wegner's argument seriously, even if it fails as a wholesale denial of free will.

[45] This, as we shall see below, is an important point: inference to the best explanation violates what in logic is known as "monotonicity". This means that while two explanations might equally well explain a limited set of data, the introduction of additional data might render one of the hypotheses untenable because unable to account for the new evidence. Continuing with the example from the footnote above, if a poltergeist was to appear all of a sudden, the poltergeist-explanation would now be uniquely able to explain all the observed phenomena.

I will meet this challenge with two quite different argumentative points, considering *FE** and *CEC* apart. First of all, I will show that there is a limited but significant revisionist claim to be made from Wegner's argument about *part* of our experience of conscious agency, and secondly, I will support the traditional, realist interpretation of the causal efficacy of consciousness by arguing that this remains the best explanation of the available data concerning the way consciousness is related to action initiation and control.

8.2.1 Where is the (Empirical) Will? Identifying a Successful, Limited Claim in Wegner's Argument

Wegner himself uses the observation that our experience usually corresponds to the facts of the situation as part of his argument about the role the purportedly illusory experience of conscious will can be seen to play in keeping track of which perceived acts belong to ourselves, and which are the acts of others.

One might then wonder at the use of the term "illusion", which indeed does seem altogether too strong a word for the discrepancy in question, something Wegner himself admits:

> Calling this an illusion may be a bit strong, and it might be more appropriate to think of this as a construction or fabrication. (2002, p. 2, footnote)

> Before ICW was written, I had alternate titles for it: The Construction of Conscious Will, The Experience of Conscious Will, The Fabrication of Conscious Agency, and so on. (2004, commentary p. 34, R2)

So why did he end up using 'illusion' anyway? In the same footnote, Wegner goes on to defend its usage thus:

> [T]he term illusion does convey the possibility that we place an erroneously large emphasis on how will appears to us and assume that this appearance is a deep insight. (2002, p. 2, footnote)

This points to an element of Wegner's argument that I think has sometimes been overlooked by critics responding to it – although, to the readers' defence, it seems Wegner is confused about this himself. Early on in the book, Wegner distinguishes between two commonsense notions of will: The first is will as an experience associated with certain acts, a "feeling of voluntariness". The second notion is of "conscious will as a force of mind, a name for the causal link between our minds and our actions." (2002, p. 3) After first examining will as an experience, however, the second notion of conscious will as force of mind – "will power" (2002, p. 14) – is also examined and explained wholly through various interpretative mechanisms, ultimately captured in the theory of apparent mental

causation. It is another mere appearance. The thing studied is thus always the "phenomenal will", never the "empirical will", whatever that might be. However, given that "empirical will" is defined as the actual causal relation between thoughts/persons and actions, it is arguably the "empirical will" and whatever role consciousness plays in *this* that is the interesting question when discussing the Dilemma and the possible threat to free will from epiphenomenalism of consciousness.

While Wegner to some extent vacillates between concluding that consciousness is an epiphenomenon with regards to action, and admitting it an important (although unconventional) role in this, it is, I believe, and taking the argument as a whole, reasonable to conclude that he intends to deny the common understanding of *CEC*, i.e. deny that consciousness or conscious will plays any direct role in bringing about action. Thus, the fact that he only discusses the role of consciousness in phenomenal will reflects the fact that this is the only role he takes consciousness to play. For Wegner, consciousness does not partake in the empirical will at all – at least not in any direct way commensurable with what he takes to be our commonsense or traditional ideas about freedom and responsibility.

I have already shown why this expansive claim about the role of consciousness in all aspects of action generation/initiation and control lacks conclusive support in the evidence presented by Wegner, but given the distinction between the possible roles of consciousness in phenomenal and empirical will, we can now identify a second, more limited claim from Wegner's argument, namely a claim about how an important aspect of our experience of agency is constructed based on interpretation or inference.

Given such a separation, one can both admit that the evidence Wegner brings together makes a strong case for the more limited hypothesis, and hold that no part of the evidence excludes the possibility that consciousness plays an active, direct role in empirical will – i.e. *in addition* to the experience Wegner identifies as interpretational. In terms of the Dilemma and *FE*, *FE*/CEC*, the limited claim invites an addendum to *FE**, but does not licence any conclusion about *CEC* or the veracity of *FE** in general, nor does it undermine *FE*. The addendum, which following Wegner may be called the "authorship emotion" *AE*, can be included into *FE** to make a revised version *FE†*:

> *FE†* We have the experience of consciously initiating and controlling action.
>
> > *AE* Part of this experience is the result of interpretation, and may therefore be misleading in comparison with objective, third person accounts.

Accepting the limited claim has the benefit of explaining the exceptions of automatisms and illusions of control, i.e. why we can be mistaken in ascribing agency, even when the agent in question is us: we can be so mistaken because an important element of our judgements (whether implicit or explicit) is the result of inferential processes which may err either from imperfect data or simply by "procedural" error. The limited claim gets us this without committing to epiphenomenalism about consciousness in action, i.e. without also denying *CEC*. Because of this, it may also be incorporated into the traditional interpretation, and the traditional interpretation, bolstered by the limited claim, will now let us account both for the standard case of correspondence between first and third-person accounts of agency, and for the exceptions of automatisms and illusions of control – without (in contrast to the revisionist interpretation) having to subsume either under the other. The *lacuna* is filled in.

Of course, if Wegner's argument were to count as a successful denial of free will according to the framework of the debate that has been set up here, he would need to show that *CEC* is false, i.e. that consciousness in not causally effective in initiating and controlling action. As I have argued, the evidence presented by Wegner is not sufficient to deny *CEC*, and so his argument does not constitute a successful empirical denial of free will. Even so, it might still be the case that *CEC is* false, and that this explains the evidence presented by Wegner – in which case his argument would regain status as a contender for an inference to the best explanation, even tipping the scales in favour of the revisionist interpretation; since if *CEC* is false, what we now take to be the ordinary case must involve some kind of illusion or mistake, e.g. the one associated with Wegner's expansive claim. Settling the question of whether *CEC* or something like it is true of actual humans is therefore a prime concern in adjudicating the debate between the traditional and the revisionist interpretation. Additionally, if *CEC* is one of the true dependencies of our traditional or commonsense ideas about free and responsible action, investigating whether it or something like it holds for humans is also directly relevant to the question of free will. Onwards to the causal efficacy of consciousness, then.

9 Countering Epiphenomenalism

> "It's not as though the task of neuroscientists who work on free will has to be to show there isn't any." Alfred Mele (Smith 2011)

9.1 Substantiating the Causal Efficacy of Consciousness

Having identified the (arguably successful) limited claim in Wegner's argument, and having incorporated *AE* into *FE†*, we can move on to investigate whether there is evidence to support *CEC* beyond intuitive plausibility. This will provide support for the interpretation that what remains of *FE** in *FE†* (which is most of it) is a veridical experience, what in turn, and together with the refutation of arguments for *FT* in Part 1, can be taken as support for the claim that *FE* is a veridical experience, and hence endorse a continued belief in free will.

The first thing to be noted here is that the main source of support for *CEC* just is its intuitive plausibility based on everyday experience. I have already argued that we experience consciously initiating and controlling our actions (*FE**), and I take this to be a fairly unproblematic statement in itself; all the more so with the inclusion of the "scientifically vetted" addendum *AE* (in *FE†*). The revisionists also appear to be interested only in undermining the basis for *FE**, and have not questioned the assumption – that *FE** is true of humans – itself. In the following discussion, I will therefore continue the dialectic between a traditional interpretation that is realist about *CEC*, and the competing revisionist interpretation that continues to maintain that consciousness is causally inefficacious with regard to action. In order to support *CEC*, I will present both theoretical considerations and empirical situations for and in which the traditional interpretation exceeds the revisionist in its ability to explain the matters at hand. While this cannot conclusively establish *CEC*, it will render the revisionist interpretation so contorted that it can be disregarded in practice.

The first step towards this goal is to ask what exactly *CEC* should be taken to entail, i.e. what role or roles consciousness can be taken to play with regard to action.

9.2 What is Conscious Will (for)?

While Wegner is not arguing that the experience of conscious will is strictly epiphenomenal, he does argue that it is not one of the causes of action in any direct way corresponding to traditional ideas about how conscious decisions issue in action. There is, in other words, a kind of limited or local epiphenomenalism in Wegner's (expansive) claim, an epiphenomenalism about direct conscious involvement in behaviour generation – approximately what Nahmias calls "modular epiphenomenalism" (Nahmias 2002, p. 530): Wegner seems to regard consciousness or conscious

will as a separable module which receives input from causally effective brain processes, but which has no output back into those processes.

This way of framing consciousness or conscious will brings me back to the first issue identified in the assumptions of Libet's original argument, namely the open question of whether this particular conception of how and when consciousness can be said to impact on behaviour is the best or most appropriate. Seen in terms of *CEC*, we should consider at what level the hypothesised "[conscious] initiation and control of action" is supposed to be taking place.

9.2.1 Recasting Consciousness

The idea that consciousness does not matter to action is an old one in psychology and philosophy. Shaun Gallagher, taking a cue from William James (1842-1910), credits the comparatively unknown British philosopher Shadworth Holloway Hodgson (1832-1912) – or, rather funnily, "Hodgson's brain" – with explicating the idea that "[n]eural events form an autonomous causal chain that is independent of any accompanying conscious mental states." (Gallagher 2006, p. 109) According to the tradition Hodgson represents, "[c]onsciousness is epiphenomenal, incapable of having any effect on the nervous system." (Gallagher 2006, p. 109)

Gallagher also quotes James in saying that Hodgson's claim is the natural completion of Descartes' argument about animals being mere automata. Descartes famously stopped short of claiming the same of humans, attempting instead to argue that we have an uncaused immaterial soul that causally interacts with the brain through the pineal gland.[46] The problem with Descartes' attempt to save a role for the soul in his otherwise completely mechanistic picture was to account for how an immaterial entity could interact with a material body. In his *Treatise of Man* (1662/1664), he simply specifies that one of the ways movement of the muscles can be brought about is by "the force of the soul" acting on the gland (Lokhorst 2011). Hodgson's solution to this proto-form of agent causation was simply to extend the evident explanatory power of an autonomous nervous system to cover human action as well. Presumably, he could not deny the self-evident fact that there was something like Descartes soul, i.e. a consciousness, but this move obviated the need to account for the interaction of mysterious soul and physical body, since there on this account simply was no such interaction.[47]

[46] The only brain structure known at that time of which there was only one, in contrast with the other features which are mirrored in the opposing hemispheres; a fact which for Descartes explained the unity of perception/consciousness (Lokhorst 2011)

[47] This kind of epiphenomenalism has its own *lacuna* however, since it is hard to understand how the immaterial soul is kept "in tune" with the material brain if there is no interaction in either direction.

According to Gallagher, the uneasy Cartesian middle position continues to shape the debate concerning what role (if any) consciousness can be said to play in relation to action initiation and control: Consciousness is seen as a kind of mind space (the "module" in modular epiphenomenalism) in which action can be simulated and from which consciousness can or cannot (as the revisionist claim) direct purely physical/non-conscious bodily movement. When this is taken as a premise in the debate concerning free will, the proof of free will ends up hinging upon whether or not one can show consciousness *in action*, directly initiating and controlling the execution of motor acts (Gallagher 2006, pp. 110ff).

Experiments like those by Libet follow this train of thought, looking for proof of *CEC* in the immediate vicinity of action – both in terms of temporal order and in the kind of relation between consciousness and motor acts that is taken as paradigmatic. Thus, W – which is supposed to be a "spontaneous intention" to move immediately preceding the execution of the simple motor act of flexing a wrist or a finger – is taken to be an appropriate measure of the role consciousness plays in action initiation. The temporal aspect in particular is weighted, with *CEC* – and by explicit extension, conscious will – either being killed off by concluding that consciousness comes too late (part of the argument for Wegner's illusion), or also (purportedly) salvaged by appeal to some last-millisecond intervention into the execution of unconsciously initiated motor plans (Libet's veto).

I have already shown why (extant) Libet-style experiments do not allow one to conclude anything about *CEC*, and this point can be expanded by the recognition that the framing of consciousness in these experiments only captures a small subset of the theoretically possible ways in which consciousness could be causally effective with respects to behaviour. I do not wish to conclude that consciousness does not have a role to play in the direct and immediate initiation of (simple) motor acts (indeed, I believe it has, and evidence for this will be reviewed below), but it is clear that such a framing misses important aspects of what it means for us *consciously* to do something. To see this, it suffices to recognise that things like conscious deliberation and choice are usually not explicitly related to the execution of motor acts, nor temporally related to these on the scale of milliseconds. One response to the results from Libet-style experiments has therefore been to focus instead, e.g. on the conscious, explicit choices that participants make about partaking in the experiments, and the intent they have formed of following the experimenter's instructions. For these choices, which undeniably are followed by the behaviour thus chosen, no neuroscientific evidence is provided to cast doubt on their causal efficacy.

In line with this, Gallagher argues that the common framing of the question "does consciousness cause behaviour?" in terms of movement initiation and control is misguided, at least when it comes to deciding whether something like conscious will is operative in a way that is relevant to freedom and responsibility. According to him, the paradigmatic conscious control necessary for the exercise of (free) will is not to be found in the moments before or during motor action, but in the consciously made decision to do something, i.e. in relation to actions that are intended (Gallagher 2006, pp. 117ff). As an example of this alternate way of framing "free will" in terms of consciously chosen intentional action, Gallagher compares the lightning fast reaction he might have to a something moving in the grass next to his feet (at time *T*), where he jumps back before having had time to realise what he has encountered – behaviour which can be explained wholly through non-conscious processes –, to the subsequent behaviour he might consciously bring about based on his wish to capture what turns out to be a harmless lizard for his collection:

> My next move is not of the same sort [as the initial reflex]. At T+ 5,000 ms, after observing the kind of lizard it is, I decide to catch it for my lizard collection. […] At *T* + 5,150 ms I take a step back and reach for [it]. One could focus on this movement and say that at T + 4,650 ms, without my awareness, processes in my brain were already underway to prepare for my reaching action, before I had even decided to catch the lizard – therefore, what seemed to be my free decision was actually predetermined by my brain. But this ignores the context defined by the larger time frame, which involves previous movement and a conscious recognition of the lizard. (Gallagher 2006, p. 118).

Gallagher argues that his reaching for the lizard cannot be explained without referring to the intention he has of catching it, which in turn refers to his conscious decision to do so and the matrix of judgements and goals that surround this decision.

Considered in the terms of competing explanations that we have used so far, we could certainly imagine a revisionist interpretation of the above scenario. However, this would either involve a duplication of the proposed causally efficacious conscious entities (e.g. his decision and intention) into unconscious ones of comparable function – at which point one would be hard pressed to find a reason to choose these hypothesised non-conscious copy-entities over their extant conscious originals –, or one would have to provide an explanatorily superior account of Gallagher's behaviour without making any reference to conscious entities or words like 'reason', 'goal', 'decision', etc. While this might be possible, the onus is on the revisionist for providing such an explanation. Provided that no such explanation is (immediately) forthcoming, we are permitted to conclude that the traditional interpretation is the best explanation available.

Note also that this does not entail any form of substance dualism: Gallagher's conscious decision to catch the lizard is not some extra entity over and above the workings of his brain. The claim is simply this: given the facts of the matter, the observable difference between his initial withdrawal from the lizard and his subsequent reaching for it is best explained by the fact that he in the second case has formed the intention to catch the lizard. If there is to be talk of any causal relations in this explanation, it must therefore also be between the relevant conscious phenomena and the subsequent behaviour.

In order to bolster the claim that certain actions can only be explained by reference to conscious decisions, and to clarify how this supports *CEC*, we will now turn to Alfred Mele's work on the role of consciously formed intentions to act in understanding conscious behaviour.

9.2.2 Effective Intentions

The subtitle is taken from a book by Alfred Mele (2009), who approaches the theme of consciousness and free will with the notions of "intention" and "conscious decision". Mele argues that the meaning of these two words are rarely if ever made explicit in research on consciousness and behaviour, and that a proper understanding of them in relation to what he terms "practical deciding", i.e. deciding what to do, can throw light on the role played by consciousness with regards to action. While Mele's work as a whole provides a useful framework within which to discuss free will from a scientific perspective, the current focus is on his argument for how conscious decisions are causally efficacious (Mele 2010).

As Mele conceives of it, (consciously) deciding to do a thing *A* is to perform a *momentary* (mental) act of forming the intention to *A*.[48] Importantly, the decision must not be confused with any process that might lead up to it, e.g. deliberation about what to do. A decision is thus an "intention-forming action" (Mele 2010, p. 3), circumscribed as the act itself. There are *proximal decisions* concerned with the immediate performance of some act, and *distal decisions* that pertain to an act to be performed at some future time (and, correspondingly, proximal and distal intentions formed through such acts of deciding).

Proximal decisions are what Libet-style experiments are concerned with (even if the original Libet *et al.* experiments didn't test decisions, as discussed above), and Mele confesses to taking part in a Libet-style experiment where he failed to feel any "urge" to move at all. As a way around this (it would arguably be embarrassing if he sat motionless throughout), he devised a way of immediately

[48] More precisely, "an executive assent to a first-person plan of action.", where "executive assent" is what you do when you consciously assent to some suggestion or proposition (Mele 2010, p. 44).

deciding to move by silently saying "now!" to himself as the triggering of finger movement by conscious choice. Mele argues that there is no way in which consciousness can lag behind such a decision: in contrast to conscious awareness of external events – which may only arise some time after the external event is already underway[49] – the act of silently saying "now!" is not a thing of which Mele can become aware after it has already begun; the act of (mentally) saying "now!" just *is* the conscious silent "now!"-saying. Mele vehemently denies any suggestion that he is a substance dualist because of this stance, since the conscious speech act of silently saying "now!" undoubtedly is the mediate result of a causal process, and not some Cartesian "force of the soul". But the process leading up to the speech act is not the speech act itself, and so whether or not one is aware of this process is not at issue: what matters is that the conscious speech act is conscious right from the onset (Mele 2010, pp. 5-6).

The suggestion is that a conscious decision to move can be made in the same manner, i.e. as a momentary act that is conscious throughout its (short) lifespan. Indeed, silent "now!"-saying can arguably be understood as an imperative "flex now!" (Mele 2010, p. 6), i.e. something like an inner vocalisation of the conscious proximal decision to move (now!). Given that Mele hit on this particular strategy because no urge or intention to act arose on its own, the conscious decision to flex, made explicit in this paradigm by the silent "now!"-saying, directly explains his subsequent flexing; he would not have acted otherwise. While it would be possible to provide a revisionist (epiphenomenalist) explanation of how Mele's silent "now!"-saying resulted in him flexing his finger, such an explanation would be at an additional disadvantage to the traditional (realist about *CEC*) explanation compared to the example of Gallagher's intention above, since the revisionist here would also need to explain the apparent *necessity* of a conscious phenomenon for action that at the same time is causally inefficacious with regards to that action.

When considering a specific act performed at a specific time t, there is a further argument to be made for the interpretation that the conscious decision to perform the act is causally effective in bringing it about: assuming a Libet-style experiment in which subjects are asked explicitly to make a decision to flex immediately (as was the case in Pockett and Purdy 2011), and assuming that an unconscious proximal decision to flex would be as effective as a conscious proximal decision; for any given trial, the subject *Sam* performs the conscious act of forming the intention to immediately flex at time t and goes on to flex immediately at time tf (where an instance tf_i can be taken to follow an instance t_i by a certain number of milliseconds). Now, for Sam, for any given point in time during

[49] His example is of the Frölich effect, where a moving slit of light is perceived only some distance from the edge at which it first appears, never at the edge.

the experiment, t_1, t_2,…, t_n, he might perform the momentary conscious act of forming a proximal intention to flex, or not. And for any given time Sam does flex, tf_i, the act of flexing can be explained either by appeal to a non-conscious or to a conscious proximal decision to flex made at the corresponding time t_i. But, argues Mele, if Sam is good at following instructions, if he does not consciously form a proximal intention to flex at t_1, it is highly unlikely that he will go on to flex at tf_1 – rather, we would expect that Sam waits until t_2 and *then* consciously forms the proximal intention to act at that time, thus flexing his finger at tf_2. Given this setup, the fact that Sam acts *when he does* is explained by his conscious proximal decision, and so it seems that this is causally relevant to bringing about the flexing of his finger (Mele 2010, pp. 52-3). Thus, even if we grant that there is such a thing as a non-conscious decision and that this is as effective in bringing about behaviour as conscious decisions appear to be (claims that are not independently argued for by Libet and others who can be taken to assume them), it still seems that only the conscious proximal decision can explain why action has been perform *at a specific time*, and not at some other.

If we yet again consider this a standoff between competing traditional and revisionist interpretations of the data, the conscious proximal decision is arguably the most powerful of the competing explanatory entities: if it is true that there is such a thing as a *causally efficacious conscious proximal practical decision*, this would explain Sam's actions in the imagined experimental setup. If, on the other hand, no such thing exists, and Sam's actions are to be explained using something like a *causally efficacious non-conscious proximal practical decision*, it is difficult to understand why such a thing should be accompanied by what must then be understood as a *causally inefficacious conscious proximal practical decision*, and apparently necessarily so. The revisionist interpretation suffers from what appears to be a malicious multiplication of explanatory entities without any gain in explanatory power (which makes it less than an ideal contender for a claim of an inference to the best explanation). I take this to be evidence for the claim that consciousness is causally efficacious also in the immediate initiation of simple motor acts (even if the translation of a conscious decision into motor behaviour might be taken care of sub-consciously).

The point about conscious proximal decisions explaining why action is performed at a specific time can also be extended to *distal* conscious practical decisions, i.e. the momentary formation of an intention to perform an act at some specific later time, which is particularly relevant to the conception of "conscious will" that Gallagher argued for above. In this regard, Mele points to the work done by Peter Gollwitzer on so-called "implementation intentions" (Mele 2010, pp. 54ff). An implementation intention is, simply put, a specification of where, when and how some already-set goal is to be achieved. Research shows that forming explicit implementation intentions drastically

increase the rate of goal attainment for people who are otherwise similarly motivated to attain the same goal, e.g. for two groups of women who decide to do a breast self-examination (BSE) sometime during the next week: In the control group, who simply decided to perform the BSE at some unspecified later time that week, 53% of the women followed up on this intention. In the test group, who in addition to this made a written statement specifying where and when they would perform the exam, 100% of the women performed the BSE. Similar results were obtained both for compliance with an exercise regime (compliance rose from 29 to 91%) and with drug addicts who were to write a CV (0 vs. 80%). In parallel with the argument about Sam's proximal conscious decision to act at time t_i being the best explanation for why he flexed his finger precisely at tf_i, Mele argues that the high goal attainment rates of the groups who were told to form explicit implementation intentions is best explained simply by reference to the fact that they formed explicit implementation intentions. While the recourse to a revisionist explanation through unconscious processes is still available, this would now not only have to account for a (by now familiar) necessary-but-epiphenomenal entity, here the conscious implementation intention; but would also have to provide some non-conscious mechanism which would explain how the non-conscious formation of an implementation intention would carry over to affect goal attainment days or even weeks later, without relying on the proposed causally effective conscious processes (Mele 2010, p. 57).

I am, of course, still relying on a glorified form of intuitive plausibility, but given that the revisionist (or illusionist as Mele calls him) has not provided any good reason to doubt the causal efficacy of conscious decisions, intuitive plausibility counts towards the conclusion that *CEC* is true of humans.

These theoretical considerations appear to give strong, independent support for *CEC*. In keeping with the theme of this thesis, we should therefore also ask if there are scientifically studied phenomena that the traditional interpretation is uniquely able to explain. To this end, I will consider two quite different kinds of pathologies that are alike in that they change the relationship between conscious and subconscious processes for action control, and thereby reveal what appears to be clear evidence in favour of the traditionalist interpretation that consciousness is causally efficacious in initiating and controlling action.

9.3 Empirical Evidence for *CEC*

Among the things Wegner appeals to in his argument to the effect that conscious will cannot be what directly causes or initiates action, is the fact that much of what we do – both as (what might be called) "everyday experts" navigating the physical and social world with ease, and as real experts who perform amazing feats in competitive sport, music and the like – is done without any feeling of "willing", i.e. without direct conscious control over the proceedings (Wegner 2002, pp. 83-4, see also Bargh and Chartrand 1999). Indeed, at a certain point, trying to retain explicit control over the execution of bodily movement will only serve to worsen performance, as evinced by pro baseball players who suddenly "forget" how to throw a ball when self-consciousness imposes itself on the otherwise smoothly unfolding action program that constitutes a skilled throw (what in sports is known as "choking", Demak 1991). But while these considerations can perhaps be taken as support for the revisionist interpretation if coupled with more stringent arguments for this, it is certainly also possible to give them a traditionalist spin, e.g. if we, keeping in mind Gallagher's argument above, focus on the amount of training that has gone into achieving a level of expertise where complex feats can be performed without direct conscious control (or, simply, remark that consciousness does *something* to action if it can disrupt it).

That training, as opposed to (some forms of) skilled execution, involves direct conscious control is here (yet again) taken as *prima facie* plausible fact, and while I have yet to see a convincing argument against this, it behoves the traditionalist interpretation to produce some additional, empirical evidence for *CEC*. However, instead of trying to substantiate the possible role of consciousness in practice and training (what I imagine to be a complex endeavour, considering the myriad possibilities for revisionist objections), I will focus instead on two quite distinct pathologies where consciousness appears beyond doubt to be directly responsible for action initiation and control.

9.4 The Role of Consciousness in Deafferentation

As with so many of our bodies' amazing yet transparent abilities, the importance of subconscious motor control is perhaps only evident when something goes amiss. Persons having suffered complete deafferentation provide one grimly fascinating example of this. The most famous case of deafferentation is that of Ian Waterman, known in the literature as "IW", who at the age of 19, after what probably was an autoimmune reaction, lost all his peripheral sensory nerves below the neck (Cole 1995). IW was not paralyzed in the traditional sense, since all nerves going from the brain out to muscles were still intact, but without any "proprioceptive feedback" providing him with information about the location and movement of his torso and limbs, he was unable to control any movement. Initially slumped in a chair or lying secured on a bed with complete lack of control over

his body, IW followed a rigorous training regime (of his own making) in which he would endlessly repeat different component of movements, gradually building up mastery of his body by checking *visually* the effects of a given effort to move, and then modulating the effort based on the effect he could see. IW has in effect created an external, visually based feedback loop to make up for the loss of feedback from proprioception (McNeill, Quaeghebeur, and Duncan 2010).

Of course, even if the deficit can be made up for by consciously attending in this way, it is still a tremendously disabling condition because of the sheer amount of waking consciousness that has to be dedicated to monitoring movement (IW's life is, as the title of Cole's book reflects, "a daily marathon"). And so, by the misfortune of others, the blessing that is subconscious motor control is evident for the rest of us: we are able to think, talk, and generally multitask while our bodies take care of themselves; we only have to specify where we want to go, and we will usually know how to get there without having to think of how to move our legs. In other words, in the normal situation, most of our voluntary movements are directly executed and modulated or controlled by subconscious processes. Therefore, even if we do not find direct, conscious control over such movements in healthy subjects, we should not be surprised: this is not where conscious control is usually exercised. Seen this way, the case of IW seems to point directly to the causal efficacy of consciousness, namely in the successful substitution that takes place between subconscious and explicitly conscious control processes.

9.4.1 An Objection from Epiphenomenalism about Visual Consciousness, Refuted by the Example of Blindsight

The revisionist advocate might object to this presentation, and argue that it is not consciousness or conscious attention *per se* that allows deafferentated subjects like IW to guide their movements, but rather the visual feedback loop itself, regardless of any conscious awareness of what enters into it. The accompanying phenomenology, the revisionist could argue, is no more causally effective here than in the experience of conscious will; accurate movement simply depends on information being passed through the visual system, and because the visual system and consciousness are so closely related, this particular (epiphenomenal) phenomenology naturally accompanies it.

Somewhat surprisingly, the idea that visual information guides action independently of any consciousness of that information appears to be borne out in part by one widely influential view of higher order visual processing in the brain, the two-streams hypothesis (Goodale and Milner 1992, Milner and Goodale 2006). According to the two-stream hypothesis, visual information is processed in two functionally independent and anatomically distinct "streams", the *dorsal* and

ventral streams. Crudely put, the ventral stream is the "what"-stream, tasked with object recognition and other elements of vision that contribute directly to our conscious experience of seeing (i.e. "perception"). The dorsal stream, on the other hand, is the "how"-stream, responsible for interpreting visual stimuli geared towards action, e.g. providing visual information to the subconscious processes that modulate posture and movement in relation to the environment. There is also empirical evidence for a double dissociation between the two streams: It has been shown 1) that visual information can successfully guide a subject's action towards an object that the subject in question is unable to identify visually due to lesions on the ventral stream. This is complemented 2) by the finding that subjects, due to damage to the dorsal stream, may be unable to act towards an object that is clearly visible and that they can successfully identified (Milner and Goodale 2006). One might therefore imagine the revisionist finding support for his interpretation by claiming that the substitution of visual for proprioceptive feedback in IW happens entirely through the non-conscious processing of visual information in the dorsal stream. And because IW has normal vision, any visual input will *also* be processed in the ventral stream, thus giving rise to the conscious visual experience, which on the revisionist interpretation is epiphenomenal (with regard to action).

However, the revisionist is not thus supported by the two-streams hypothesis: even if the two streams may be functionally and pathologically *dissociable*, they are not dissociated in the normal, non-pathological case. So it does not make good sense to say that IW is *either* consciously *or* subconsciously using the visual information to help control the movement of his limbs: as long as his vision and higher order visual processing is normal, he is engaged in both at the same time, and could not will to do otherwise.[50] Furthermore, there are also questions as to whether the two streams really are functionally independent, or if the differences found are better explained by a relative functional specialisation of associated brain areas (Schenk and McIntosh 2009). In that case, the potential support for the revisionist interpretation would be further weakened.

The revisionist might object again: there are cases where people have lost either one or both of their primary visual cortices (V1), people who appear to be partially or completely blind, who can still act on visual information present in their blind field – what is known in the literature as "blindsight". Do these cases not show that subconscious processes are self-sufficient for action initiation and control, and that consciousness is unnecessary? The revisionist might here point, e.g. to patient TN, who after a stroke destroying his entire V1 normally relies on a white cane to get around. However, in the hands of Beatrice de Gelder and Lawrence Weisenkrantz (the latter of whom coined the

[50] For a recent review of the complex interactions and top-down effects of the visual processing system, see (Gilbert and Li 2013).

term "blindsight", seeWeiskrantz et al. 1974), TN could easily navigate an obstacle-strewn hallway without any aid – and without having any reportable awareness either of the obstacles or his deft avoidance of them (de Gelder 2010).[51]

While this remarkable feat does indicate that the relationship between awareness of visual information and the ability to utilize visual information for action is more complex than we might previously have thought, it still does not supply the revisionist with the kind of support he needs; blindsight actually provides an excellent *counterexample* to the epiphenomenalist view of consciousness in action, since, as Antti Revonsuo notes:

> [...] the patients cannot really use the nonconscious information to any useful purpose in their deliberate behaviour or decision making. They do not know directly about the existence of that information, thus they do not see or recognize the stimuli as far as they are concerned, *and their behaviour in everyday situations is as helpless as that of a person who has neither the conscious nor the nonconscious information available!* (Revonsuo 2010, p. 130, emphasis mine)

Similarly, de Gelder notes of her patient TN:

> […] TN views himself as a blind person, and he will remain totally dependent on his white cane until he is convinced he can see without knowing it. (de Gelder 2010)

In other words, "blindsight" only appears as a phenomenon in the force-choice situation of the laboratory setting, and here the experimenter and her accomplices effectively supply the conscious planning and prompting necessary to complete action. The visual information is "there" – probably mediated by small structure known as the *superior colliculus* (SC, de Gelder 2010) – but without entering consciousness, it might as well not have been, since the person concerned cannot *choose* to act on it. Outside of the laboratory, people with blindsight are, to all intents and purposes, simply blind (or blind in part of their visual field).

The comment de Gelder makes about convincing TN that he can "see" without conscious awareness of that which he sees is also interesting in this respect: de Gelder hopes that training blindsighted subjects like TN in relying on their subconsciously present visual information might aid them in navigating everyday situations. The fact that TN must be convinced, i.e. be persuaded to consciously embrace a belief in his own, subconscious powers of vision for this to be of any use to him, is a striking example of the role of consciousness in turning information into action.

---

[51] There is a video of this available at the Scientific American *Observation* blog web pages, attached to the article "Blindsight: Seeing without knowing it" (Collins 2010). There is a video of this available at the Scientific American *Observation* blog web pages, attached to the article "Blindsight: Seeing without knowing it" (Collins 2010).

While this does not prove the causal efficacy of consciousness beyond any sceptical epiphenomenalist doubt, it does make the revisionist interpretation even more contorted, since the *de facto* blindness of blindsighted subjects means that the revisionist must account for a consciousness to which (visual) information necessarily must be presented in order for it to initiate action[52], but which has no (direct) causal influence on the action thus initiated. The traditional interpretation of the data has no such problems, and can easily incorporate the dual findings of the power of subconscious processes in guiding what at some level or point must be consciously chosen action.

❃

Overall, I take the above theoretical considerations and empirical evidence to provide strong support for the traditional interpretation that *CEC* is true of humans. There remains, to my knowledge, but one possible objection from the revisionist before I can conclude the argument of this thesis.

9.5 The Objection from Temporal Priority Again…

Granted that my argument thus far is successful in establishing a causally efficacious role for consciousness with regard to action, the revisionist might still make a more fundamental objection to the claim that humans have something like the empirical free will we are discussing here. The objection goes something like this:

Even if consciousness *is* causally effective with regard to behaviour, e.g. when we consciously decide to do something, and even if RP is not the "real cause" of voluntary actions, conscious decisions are nevertheless the result of preceding neuronal activity. Thus, consciousness in the role of making decisions can only be seen as the *mediate* cause of behaviour, i.e. a part of a causal chain. And in this causal chain, the arrow of causation still goes from sub- or non-conscious to conscious processes (even as it might go on to behaviour from there). In other words, even with the expanded time frame as argued for above, consciousness still always "comes too late" to fulfil the role ascribed to it in traditional or commonsense conceptions of free will, because it is not the first this in a chain of events.

The objection is nevertheless easily refuted, since as long as consciousness is granted causal efficacy – whether it is as mediate, immediate or some other kind of cause – *CEC* is true, and the empirical

[52] Or, as Gelder suggests, the lack of such direct availability may be made up for by training oneself to consciously rely on non-conscious information.

denial of free will by way of arguing for epiphenomenalism has failed. If there is a lingering feeling of unease with this resolution, I think I know the source: Admitting that our conscious decision to move is (merely) a mediate and not an *ultimate* cause activates underlying incompatibilist intuitions, as discussed in Part I. However, as we also saw in Part I, this appeal to ultimate causes is not straightforwardly relevant to the question of whether or not we have free will; specifically, there is no reason to accept the claim that having free will necessarily entails being the (conscious) ultimate cause of one's actions.

Another way of approaching the objection from temporal priority is to say that it simply does not make sense to claim that consciousness "comes too late" due to preceding brain activity: the brain is continually active (unless one is brain-dead), and, still barring substance dualism, consciousness is somehow part or an aspect of this constant activity. We can thus *always* find brain activity preceding a conscious phenomenon if that is what we are looking for, but given that no one has found a "neural correlate" for consciousness, nor even an unequivocal correlation between patterns of neuronal firing and a single conscious event (to the extent that such an event can be meaningfully isolated from the stream of events), this only amounts to a futile exercise in generating ever new, ever question begging statements of correlation.

Finally, the entire worry from epiphenomenalism due to temporal sequence reveals a strong bias towards thinking of consciousness and behaviour in terms of *linear causality*, and it is worth, as a final side-note, to question whether this is suitable for describing the constant, often circular activity of the human organism with its brain and attendant consciousness.

9.5.1 … And a Short, Additional Reply Using Circular Causality

Walter J. Freeman (2006)[53] argues that linear causality, while useful for describing technological relations and as shorthand for the workings of medical interventions, is too blunt a theoretical tool to be used to understand the complex, nonlinear interactions of our brain and mind. While we might describe the operation of a machine, or even certain kinds of human behaviour at a sufficiently abstract and circumscribed level using clearly defined "start" and "stop" points; *stimulus* and *response* conditions, this is inappropriate when considering the human organism as it is, constantly active and constantly interacting with its environment. Freeman is (an early) member of a gradually rising number of thinkers on cognition, perception and behaviour who apply the mathematical tools of *dynamical systems theory* (DST) in order to model complicated

[53] The III, son of the (in)famous physician Walter J. Freeman II, who is known for popularising (and avidly performing) a simplified procedure for lobotomy.

cognitive/neural systems. With roots in John Dewey (1914) and Maurice Merleau Ponty (1961), and importantly developed by James Gibson (1979), contemporary scientist and philosophers like Hermann Haken, J. A. Scott Kelso and H. Bunz ("HKB", 1985); Michael Turvey (1990); Andy Clark (1998); Tony Chemero (2009) and Michael Silberstein (Chemero and Silberstein 2011) (to name but a few); argue for the view that the human organism and its environment – or rather, the "niche" created by the mutual interactions of an organism with specific demands and abilities and the environment that both shapes and is shaped by this organism – must be understood as a complex whole, a system dominated by interactions characterised by the dizzying fact that each "part" always both changes and is changed. This also means that no component can be singled out as uniquely changing, and another as being changed: influences are reciprocal and non-linear. While this can give rise to (apparent) complete chaos (which is why one also talks of "chaos theory"), such systems can also (spontaneously) organise into (temporarily) table patterns, out of which can emerge new (higher order or "emergent") properties that arguably only belong to the system as a whole. Using DST, the state of the system as it develops over time is described as moving through a "state space" of possible configurations, with "attractors" arising out of the chaos to ensnare, more or less temporarily, the changes through which the systems moves over time – reducing, but not to unity, the set of configurations within which the system is likely to develop, so that properties may arise with temporary stability without the system ever returning to the exact same configuration.

All this is terribly exotic compared to the trusty old conceptual framework supplied by linear causality, but has arisen as a direct reaction to the failures of the traditional mind-set to provide convincing models for understanding the mind, brain, organism and environment as these (undeniably) interact in the real-world case of living, breathing, thinking, feeling, acting, and perceiving human beings (and other organisms). The upshot of this is that there is no room in this picture for the worry from temporal priority: the mind is continually active and developing, and nothing can be singled out from the mutually and non-linearly interacting elements that can be said to be the "real cause" of some phenomenon like a conscious decision. As Freeman suggests, we may, for some applications, speak of causality in a more traditional sense, e.g. to say that a certain virus causes a certain disease, but only if we are aware of the fact that this kind of "linear" causality is a heuristically identified set of interactions within a larger set that is essentially circular.

10 Concluding Remarks

> "Thus, neuroscience might enable us to develop a more sophisticated view of responsibility that takes into account both the cognitive demands and the control demands made by intuitive and legal notions of responsibility, and reconciles them with a scientifically informed view of the brain as a physical system that governs our actions. This could result in a compatibilist theory of moral responsibility that is not predicated on paradoxical views of absence of causation or freedom from causal laws." Adina Roskies (2006a)

If my argument as presented here is granted, I take it to offer solid grounds on which to conclude that neuroscience has not settled the question of whether or not humans have anything like free will. Furthermore, and going beyond this purely negative claim: in light of the evidence presented, there is little doubt that consciousness is causally efficacious in bringing about behaviour. If, as suggested, i.a. by Shaun Gallagher, the empirical question of free will is about the way we consciously choose to act, and if the notion of "conscious will" as an empirical stand-in for "free will" is cashed out in terms of something like Mele's "conscious decision"; I also take my argument to have provided good reason to be a realist about *this kind* of free will.

To addend one last conditional, and therewith to tie together the two parts of this thesis: if "this kind" of free will – i.e. the causal efficacy of conscious choice unhampered by a fundamental incompatibility between freedom of choice and the physical world – is what our notions of freedom and responsibility need for their justification (if indeed any such justification is needed beyond the actuality of our reactive attitudes), then I also believe that my argument supports a cautiously positive reply to the question "do humans have free will?"

Acknowledgements

Thanks are due to Jonathan Knowles for all his helpful suggestions and critical assessments throughout the writing of this thesis. Thanks also to NTNU for being accommodating to my various requests.

Bibliography

2002. NOU 2002: 4 Ny straffelov. Seksjon 4.2.3: Straffens begrunnelse, berettigelse og hensiktsmessighet.: Justis- og politidepartementet.

Banks, W. P., and E. A. Isham. 2009. "We Infer Rather Than Perceive the Moment We Decided to Act." *Psychological Science* no. 20:17-21.

Banks, W. P., and S. Pockett. 2007. "Benjamin Libet's Work on the Neuroscience of Free Will." In *The Blackwell companion to consciousness*, edited by Max Velmans and Susan Schneider, xviii, 744 p. Malden, MA ; Oxford: Blackwell Pub.

Bargh, John A., and Tanya L. Chartrand. 1999. "The unbearable automaticity of being." *American Psychologist* no. 54 (7):462-479. doi: 10.1037/0003-066x.54.7.462.

Barkow, Jerome H., Leda Cosmides, and John Tooby. 1992. *The Adapted mind : evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Batthyány, Alexander. 2009. "Mental Causation and Free Will after Libet and Soon: Rclaiming Conscious Agency." In *Irreducibly Conscious. Selected Papers on Consciousness*, edited by Alexander Batthyany und Avshalom Elitzur, p. 135ff. Heidelberg: Universitätsverlag Winter.

Bayne, Tim. 2011. "Libet and the Case for Free Will Scepticism." In *Free will and modern science*, edited by Richard Swinburne. Oxford University Press.

Boswell, James. [1791] 2012. *James boswell's life of johnson : an edition of the original manuscript in four volumes. volume 3: 1776-1780*, *Yale editions of the private papers of james boswell*. New Haven, CT: Yale University Press.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafo. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nat Rev Neurosci* no. 14 (5):365-376. doi: 10.1038/nrn3475.

Chemero, Anthony. 2009. *Radical embodied cognitive science*. Cambridge, Mass.: MIT Press.

Chemero, Anthony, and Michael Silberstein. 2011. "Dynamics, Agency and Intentional Action." *Humana.Mente* no. 15:20.

Chisholm, R. M. 2003. "Human Freedom and the Self." In *Free will*, edited by Gary Watson, 24-35. Oxford ; New York: Oxford University Press.

Clark, Andy. 1998. *Being there: Putting brain, body, and world together again*: The MIT Press.

Clarke, Randolph. 2009. Incompatibilist (Nondeterministic) Theories of Free Will. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Cole, Jonathan. 1995. *Pride and a daily marathon*. 1st MIT Press ed. Cambridge, Mass.: MIT Press.

Collins, Graham. 2013. *Blindsight: Seeing without knowing it*. Scientific American 2010 [cited 15.04 2013]. Available from http://blogs.scientificamerican.com/observations/2010/04/22/blindsight-seeing-without-knowing-it/.

Crossley, Rosemary, and Anne McDonald. 1984. *Annie's coming out*. Repr. with revisions. ed. Harmondsworth: Penguin.

Dawkins, Richard. 1976. *The selfish gene*. Oxford: Oxford University Press.

de Gelder, B. 2010. "Uncanny SIGHT in the BLIND." *Scientific American* no. 302 (5):60-65. doi: DOI 10.1038/scientificamerican0510-60.

de Melo-Martín, I. 2005. "Firing up the nature/nurture controversy: bioethics and genetic determinism." *Journal of Medical Ethics* no. 31 (9):526-530. doi: 10.1136/jme.2004.008417.

Deacon, Terrence William. 2012. *Incomplete nature : how mind emerged from matter*. 1st ed. New York: W.W. Norton & Co.

Deecke, Lüder. 1965. *Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen : Bereitschaftspotential und reafferente Potentiale*. Freiburg (Breisgau), Universi*tat, Diss , 1965, s.n., S.l.

Demak, Richard. 1991. Mysterious Malady: Why do some major leaguers suddenly forget how to throw a baseball? *Sports Illustrated*, April 08.

Dennett, Daniel Clement. 2003. *Freedom evolves*. New York: Viking.

Dewey, John. 1914. "Psychological doctrine and philosophical teaching." *The Journal of Philosophy, Psychology and Scientific Methods* no. 11 (19):505-511.

Doris, J. M. 1998. "Persons, situations, and virtue ethics (Moral psychology)." *Nous* no. 32 (4):504-530.

Douven, Igor. 2011. Abduction. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Dowe, Phil. 2008. Causal Processes. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Dudau, R. 2002. *The Realism/Antirealism Debate in the Philosophy of Science*. Dissertation, Geisteswissenschaftliche Sektion, Universität Konstanz, Konstanz.

Eshleman, Andrew. 2009. Moral Responsibility. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Faye, Jan. 2010. Backward Causation. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Francis, Matthew. 2012. *Finding a direction of time in exotic particle transformations*. Condé Nast, 19.11.2012 2012 [cited 22.11 2012]. Available from http://arstechnica.com/science/2012/11/finding-a-direction-of-time-in-exotic-particle-transformations/.

Frankfurt, HarryG. 1988. "Freedom of the Will and the Concept of a Person." In *What Is a Person?*, edited by MichaelF Goodman, 127-144. Humana Press.

Freeman, Walter J. 2006. "Consciousness, Intentionality, and Causality." In *Does consciousness cause behavior?*, edited by Susan Pockett, William P. Banks and Shaun Gallagher, vi, 364 p. Cambridge, Mass.: MIT Press.

Gallagher, Shaun. 2006. "Where's the action? Epiphenomenalism and the Problem of Free Will." In *Does consciousness cause behavior?*, edited by Susan Pockett, William P. Banks and Shaun Gallagher, vi, 364 p. Cambridge, Mass.: MIT Press.

Gerrits, T., S. Glancy, T. S. Clement, B. Calkins, A. E. Lita, A. J. Miller, A. L. Migdall, S. W. Nam, R. P. Mirin, and E. Knill. 2010. "Generation of optical coherent-state superpositions by number-resolved photon subtraction from the squeezed vacuum." *Physical Review A* no. 82 (3). doi: Doi 10.1103/Physreva.82.031802.

Gibson, James J. 1979. *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gilbert, Charles D., and Wu Li. 2013. "Top-down influences on visual processing." *Nat Rev Neurosci* no. 14 (5):350-363. doi: 10.1038/nrn3476.

Goodale, M. A., and A. D. Milner. 1992. "Separate visual pathways for perception and action." *Trends Neurosci* no. 15 (1):20-5.

Greene, Joshua, and Jonathan Cohen. 2010. "For the Law, Neuroscience Changes Nothing and Everything." In *Neuroethics : an introduction with readings*, edited by Martha J. Farah, xv, 379 p. Cambridge, Mass.: MIT Press.

Gusnard, D. A., and M. E. Raichle. 2001. "Searching for a baseline: Functional imaging and the resting human brain." *Nature Reviews Neuroscience* no. 2 (10):685-694. doi: Doi 10.1038/35094500.

Haken, H., J. A. S. Kelso, and H. Bunz. 1985. "A Theoretical-Model of Phase-Transitions in Human Hand Movements." *Biological Cybernetics* no. 51 (5):347-356. doi: Doi 10.1007/Bf00336922.

Harman, Gilbert. 1999. "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error." *Proceedings of the Aristotelian Society* no. 99 (ArticleType: research-article / Full publication date: 1999 / Copyright © 1999 The Aristotelian Society):315-331.

Harris, Sam. 2012. *Free will.* 1st Free Press trade pbk. ed. New York: Free Press.

Haynes, J. D. 2011a. "Decoding and predicting intentions." *Ann N Y Acad Sci* no. 1224:9-21. doi: 10.1111/j.1749-6632.2011.05994.x.

Haynes, John-Dylan. 2011b. "Beyond Libet: Long-term Prediction of Free Choices from Neuroimaging Signals." In *Conscious will and responsibility*, edited by Benjamin Libet, Walter Sinnott-Armstrong and Lynn Nadel, 85-96. Oxford ; New York: Oxford University Press.

Hitchcock, Christopher. 2007. "What Russell got right." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry. Oxford University Press.

Hoefer, Carl. 2002. "Freedom from the Inside Out." *Royal Institute of Philosophy Supplement* no. 50:201-.

Hoefer, Carl. 2010. Causal Determinism. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Holton, Richard. 2013. "From Determinism to Resignation, and How to Stop It." In *Decomposing the Will*, edited by Andy Clark, Julian Kiverstein and Tillman Vierkant. Oxford University Press.

Inwagen, Peter van. 1989. "When is the Will Free?" *Philosophical Perspectives* no. 3 (ArticleType: research-article / Issue Title: Philosophy of Mind and Action Theory / Full publication date: 1989 / Copyright © 1989 Ridgeview Publishing Company):399-422. doi: 10.2307/2214275.

Inwagen, Peter van. 2000. "Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture." *Nous* no. 34 (ArticleType: research-article / Issue Title: Supplement: Philosophical

Perspectives, 14, Action and Freedom / Full publication date: 2000 / Copyright © 2000 Wiley):1-19. doi: 10.2307/2676119.

Jacobson, John W., James A. Mulick, and Allen A. Schwartz. 1995. "A history of facilitated communication: Science, pseudoscience, and antiscience science working group on facilitated communication." *American Psychologist* no. 50 (9):750-765. doi: 10.1037/0003-066X.50.9.750.

Kane, Robert. 2005. *A contemporary introduction to free will*, *Fundamentals of philosophy series*. Oxford ; New York: Oxford University Press.

Lau, H. C., R. D. Rogers, P. Haggard, and R. E. Passingham. 2004. "Attention to intention." *Science* no. 303 (5661):1208-1210. doi: DOI 10.1126/science.1090973.

Lau, H. C., R. D. Rogers, and R. E. Passingham. 2007. "Manipulating the Experienced Onset of Intention after Action Execution." *Journal of Cognitive Neuroscience* no. 19 (1):81-90.

Levy, Neil. 2007. *Neuroethics*. Cambridge, UK ; New York: Cambridge University Press.

Lewontin, R. 1982. Biological Determinism. In *The Tanner Lectures on Human Value*. Delivered at The University of Utah: Tanner Humanities Center.

Libet, B. 1998. "Do we have free will?" *Faseb Journal* no. 12 (4):A137-A137.

Libet, B., C. A. Gleason, E. W. Wright, and D. K. Pearl. 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* no. 106 (Pt 3):623-42.

Libet, B. Wright E. W., and C. A. Gleason. 1982. *Readiness-potentials Preceding Unrestricted 'Spontaneous' vs. Pre-planned Voluntary Acts*: I. Owen.

Lokhorst, Gert-Jan. 2011. Descartes and the Pineal Gland. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

McKenna, Michael. 2009. Compatibilism. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

McKenna, Michael, and Paul Russell. 2008. *Free will and reactive attitudes : perspectives on P.F. Strawson's "Freedom and resentment"*. Farnham, England ; Burlington, VT: Ashgate.

McNeill, David, Liesbet Quaeghebeur, and Susan Duncan. 2010. "IW-"The Man Who Lost His Body"." In *Handbook of Phenomenology and Cognitive Science*, 519-543. Springer.

McTaggart, J. Ellis. 1908. "The Unreality of Time." *Mind* no. XVII (4):457-474. doi: 10.1093/mind/XVII.4.457.

Mele, Alfred R. 2009. *Effective intentions : the power of conscious will*. Oxford ; New York: Oxford University Press.

Mele, Alfred R. 2010. "Conscious Deciding and the Science of Free Will." In *Free will and consciousness : how might they work?*, edited by Roy F. Baumeister, Alfred R. Mele and Kathleen D. Vohs, x, 225 p. New York: Oxford University Press.

Merleau-Ponty, Maurice. 1961. *L'Œil et l'esprit*. Paris: Gallimard.

Milner, A. D., and Melvyn A. Goodale. 2006. *The visual brain in action*. 2nd ed, *Oxford psychology series*. Oxford ; New York: Oxford University Press.

Morse, Stephen J. 2010. "Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note." In *Neuroethics: an introduction with readings*, edited by Martha J. Farah, xv, 379 p. Cambridge, Mass.: MIT Press.

Morse, W. H., and B. F. Skinner. 1957. "A second type of superstition in the pigeon." *Am J Psychol* no. 70 (2):308-11.

Mostert, MarkP. 2001. "Facilitated Communication Since 1995: A Review of Published Studies." *Journal of Autism and Developmental Disorders* no. 31 (3):287-313. doi: 10.1023/A:1010795219886.

Nahmias, Eddy. 2002. "When consciousness matters: a critical review of Daniel Wegner's The illusion of conscious will." *Philosophical Psychology* no. 15 (4):527-541.

Nussbaum, Martha Craven. 2001. *Upheavals of thought : the intelligence of emotions*. Cambridge ; New York: Cambridge University Press.

Pacherie, Elisabeth. 2006. "Toward a Dynamic Theory of Intentions." In *Does consciousness cause behavior?*, edited by Susan Pockett, William P. Banks and Shaun Gallagher, vi, 364 p. Cambridge, Mass.: MIT Press.

Pereboom, Derk. 2001. *Living without free will*, *Cambridge studies in philosophy*. Cambridge, U.K. ; New York: Cambridge University Press.

Petkov, Vesselin. 2005. Is There an Alternative to the Block Universe View?

Pike, Nelson. 1965. "Divine omniscience and voluntary action." *The Philosophical Review* no. 74 (1):27-46.

Pinker, Steven. 1994. *The language instinct*. 1st ed. New York: W. Morrow and Co.

Pockett, Susan, and Suzanne Purdy. 2011. "Are Voluntary Movements Initiated Proconsciously? The Relationships between Readiness Potentials, Urges, and Decisions." In *Conscious will and responsibility*, edited by Benjamin Libet, Walter Sinnott-Armstrong and Lynn Nadel, xvi, 261 p. Oxford ; New York: Oxford University Press.

Revonsuo, Antti. 2010. *Consciousness : the science of subjectivity*. New York, NY: Psychology Press.

Rice, Hugh. 2013. Fatalism. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Roskies, Adina. 2006a. "Neuroscientific challenges to free will and responsibility." *Trends in cognitive sciences* no. 10 (9):419-423. doi: 10.1016/j.tics.2006.07.011.

Roskies, Adina. 2006b. "Neuroscientific challenges to free will and responsibility." *Trends in cognitive sciences* no. 10:419-423.

Roskies, Adina. 2011. "Why Libet's studies don't pose a threat to free will." In *Conscious will and responsibility*, edited by Benjamin Libet, Walter Sinnott-Armstrong and Lynn Nadel, 11-23. Oxford ; New York: Oxford University Press.

Roskies, Adina L. 2012. "How does the neuroscience of decision making bear on our understanding of moral responsibility and free will?" *Current Opinion in Neurobiology* (0). doi: 10.1016/j.conb.2012.05.009.

Rudski, Jeffrey. 2001. "Competition, superstition and the illusion of control." *Current Psychology* no. 20 (1):68-84. doi: 10.1007/s12144-001-1004-5.

Russell, Bertrand. 1912. "On the Notion of Cause." *Proceedings of the Aristotelian Society* no. 13 (ArticleType: research-article / Full publication date: 1912 - 1913 / Copyright © 1912 The Aristotelian Society):1-26. doi: 10.2307/4543833.

Samet, Elizabeth D. 2010. Why Are Athletes and Soldiers So Superstitious? *New Republic*, http://www.newrepublic.com/article/77941/superstitious-athletes-soldiers-army-baseball-military-academy#.

Schaffer, Jonathan. 2008. The Metaphysics of Causation. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Schenk, Thomas, and Robert D. McIntosh. 2009. "Do we have independent visual streams for perception and action?" *Cognitive Neuroscience* no. 1 (1):52-62. doi: 10.1080/17588920903388950.

Schmid, Hans Bernhard. 2005. *Wir-Intentionalität*. Orig.-Ausg. ed, *Alber-Reihe praktische Philosophie*. Freiburg ; München: Alber.

Schmidt, R. C., C. Carello, and M. T. Turvey. 1990. "Phase-Transitions and Critical Fluctuations in the Visual Coordination of Rhythmic Movements between People." *Journal of Experimental Psychology-Human Perception and Performance* no. 16 (2):227-247. doi: Doi 10.1037//0096-1523.16.2.227.

Sinnott-Armstrong, Walter. 2011. "Lessons from Libet." In *Conscious will and responsibility*, edited by Benjamin Libet, Walter Sinnott-Armstrong and Lynn Nadel, xvi, 261 p. Oxford ; New York: Oxford University Press.

Skinner, B. F. 1948. "Superstition in the pigeon." *J Exp Psychol* no. 38 (2):168-72.

Smith, K. 2011. "Taking Aim at Free Will." *Nature* no. 477 (7362):23-25.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008a. "Supplementary Information: Unconscious determinants of free decisions in the human brain." *Nat Neurosci* no. 11 (5). doi: http://www.nature.com/neuro/journal/v11/n5/suppinfo/nn.2112_S1.html.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008b. "Unconscious determinants of free decisions in the human brain." *Nat Neurosci* no. 11 (5):543-545. doi: http://www.nature.com/neuro/journal/v11/n5/suppinfo/nn.2112_S1.html.

Stark, C. E., and L. R. Squire. 2001. "When zero is not zero: the problem of ambiguous baseline conditions in fMRI." *Proc Natl Acad Sci U S A* no. 98 (22):12760-6. doi: 10.1073/pnas.221462998.

Strawson, Galen. 2010. *Freedom and belief*. Rev. ed. Oxford ; New York: Oxford University Press.

Strawson, P. F. 1974. *Freedom and resentment, and other essays*, *University paperbacks*. London: Methuen distributed in the USA by Harper & Row Barnes & Noble Import Division.

Strawson, P. F. 2005. *Scepticism and naturalism : some varieties*. Digital edition ed. Vol. 12, *Woodbridge lectures*. London ; New York: Routledge/Taylor & Francis e-Library. Original edition, 1985.

Talmi, Deborah, and Chris D. Frith. 2011. "Neuroscience, Free Will, and Responsibility." In *Conscious will and responsibility*, edited by Benjamin Libet, Walter Sinnott-Armstrong and Lynn Nadel, 124-131. Oxford ; New York: Oxford University Press.

Taylor, Richard. 1992. *Metaphysics*. 4th ed, *Prentice-Hall foundations of philosophy series*. Englewood Cliffs, N.J.: Prentice Hall.

Thompson, S. C., W. Armstrong, and C. Thomas. 1998. "Illusions of control, underestimations, and accuracy: a control heuristic explanation." *Psychol Bull* no. 123 (2):143-61.

Van Inwagen, P. 1975. "The incompatibility of free will and determinism." *Philosophical Studies* no. 27 (3):185-199.

Van Inwagen, Peter. 1983. *An essay on free will*. Oxford, Oxfordshire: Oxford University Press.

Vihvelin, Kadri. 2011. Arguments for Incompatibilism. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Vohs, Kathleen D., and Jonathan W. Schooler. 2008. "The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating." *Psychological Science* no. 19 (1):49-54. doi: 10.1111/j.1467-9280.2008.02045.x.

Waller, Bruce N. 1990. *Freedom without responsibility*. Philadelphia: Temple University Press.

Watson, Gary. 1982. *Free will*, *Oxford readings in philosophy*. Oxford Oxfordshire ; New York: Oxford University Press.

Wegner, D. M., and T. Wheatley. 1999. "Apparent mental causation - Sources of the experience of will." *American Psychologist* no. 54 (7):480-492. doi: Doi 10.1037//0003-066x.54.7.480.

Wegner, Daniel M. 2002. *The illusion of conscious will*. Cambridge, Mass.: MIT Press.

Wegner, Daniel M. 2004. "Precis of the illusion of conscious will." *Behavioral and Brain Sciences* no. 27 (5):649-59; discussion 659-92.

Weiskrantz, L., E. K. Warrington, M. D. Sanders, and J. Marshall. 1974. "Visual Capacity in Hemianopic Field Following a Restricted Occipital Ablation." *Brain* no. 97 (Dec):709-728. doi: DOI 10.1093/brain/97.1.709.