Free Agency and its Place within Psychology

Michael M. Pitman

A thesis submitted to the Faculty of Humanities, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy by thesis in the field of Philosophy.
University of the Witwatersrand, Johannesburg, 2011

Abstract

Philosophical tradition locates questions about free will and agency within a debate characterised by deep cognitive tensions and a lingering sense of stalemate. Evaluating the most promising libertarian account of free will, due to Robert Kane, confirms the compatibilist worry that inserting indeterminism into moments of volition undermines claims of agency; while testing the prospects for compatibilism in a deterministic universe confirms the libertarian suspicion that free agency is not compatible with global determinism. An alternative setting for the exploration and defence of free agency is proposed, located closer to Psychology, and framed by the images of the Agent Automaton (AA) and the Hyper-rational, Hyper-reflective Agent (HHA). Giving psychological substance to the threat of the AA helps provoke fresh explorations and defences of a distinctively human, conscious free agency; while the evidence against, and questions about the normative desirability of our being HHAs argue against securing claims of free agency by making empirically and normatively unreasonable demands on our capacities for reflection, cool reason, and control. The project of explicating and defending a psychologically-informed conception of free agency, exploiting degrees of freedom in our imagination and externalised aspects of mind, is given positive substance and direction, including a speculative hypothesis for locating a freedom-friendly variety of indeterminism in processes of imaginative generativity.

Declaration

I declare that:

Free Agency and its Place within Psychology

is my own, unaided work and that all the sources that I have used or quoted have been indicated by means of complete references. It is being submitted for the degree of Doctor of Philosophy in the field of Philosophy at the University of the Witwatersrand. It has not been submitted before for any degree or examination at any other university.

Signed this _____day of_____2011

_____
Michael M. Pitman

Acknowledgements

I wish to express my sincere thanks and gratitude to the following people:

- My supervisor, Mark Leon, for years of philosophical mentorship, guidance, criticism, and the patience to see this project through to completion.
- David Martens, for his helpful and encouraging feedback on an early and tentative draft.
- My friend and colleague, Lucy Allais, for her invaluable input, support and advice, and her enthusiasm for this project in all moments when its author was struggling to maintain his.
- Brett Bowman, for his friendship, understanding, belief, and the most wonderful capacity to commiserate when it is needed most.
- My 'bosses' – Norman Duncan, Gill Eagle, Andrew Thatcher – who have had to patiently wait and cheer from the sidelines as I have made this testing journey in another discipline. Their help and support have made this project possible.
- Hugh Mellor and Simon Blackburn, for their conversation and philosophical inspiration in the early days of this project.
- My wonderful friends and family, for your endless love and support over the course of this project.
- My parents, Jen and Brian, for instilling in me a love of argument and questioning, and for supporting every step of my journey through and into the academy.
- My daughter Sophie, whose arrival in the world helped inspire her dad to imagine.
- Most of all, my amazing wife Jules – your love, companionship, support and understanding, your patience, and your unwavering belief in me, have given me the space and confidence to complete this project. Whatever its flaws, I dedicate it to you and our little girl.

I am also grateful for financial support received over the course of this project, including:

- Grants from the Ernest Oppenheimer Memorial Trust, Wits University Research Committee, Anderson Capelli Fund, and Trinity College, Cambridge that enabled a twelve week period as a Visiting Scholar in the Faculty of Philosophy at Cambridge University.
- The financial assistance of the National Research Foundation towards this research is hereby acknowledged. Opinions expressed in this thesis, and conclusions arrived at, are those of the author and are not necessarily to be attributed to the National Research Foundation.
- A staff bursary from the University of the Witwatersrand.
- A Faculty of Humanities Wits Enterprise Dividend Research Promotion Grant.

This project was made possible by this generous financial support.

Table of Contents

List of Figures and Tables

Chapter 1

Introduction

During my honours year in psychology, while busy completing a course on psychopathology, I once asked a Christian friend in the class how and where they thought manifestations of evil might be accommodated within the secular view of madness and mental illness provided by psychiatry and clinical psychology. I did not mean 'evil' in the sense associated with the so-called problem of evil in theology; nor did I mean by 'evil' the many and varied ways in which human beings show themselves capable of dramatic departures from ethical behaviour. I meant 'evil' in the sense in which I imagined a theist of a certain variety, believing not only in the existence of a god but also in the existence of supernatural forces of evil, would be committed to the existence and manifestations of evil in the world. (I don't recall receiving a satisfying answer.)

Years later, as a philosopher teaching in a Psychology department, I found myself wondering about the place of free will in psychology. At least from my point of view, this puzzle (unlike my puzzle about evil) was of a secular variety, free will being philosophically mysterious but not, as far as I could see, because of any supernatural dimensions to the puzzle. The parallel to the question about evil was rather one of wondering where and to what extent free will was factored in to psychological theory, whether basic (e.g. cognitive) or applied (e.g. clinical/ therapeutic). Any first year textbook might tell one that the Humanistic-Existentialist 'school' in psychology places great emphasis on human freedom, but such an answer is neither sufficiently general nor terribly illuminating. To the extent that psychology is in the business of articulating predictive and explanatory theories of human behaviour, in what ways (if at all) are notions of human freedom built in?

At the same time, my interdisciplinary position encouraged asking questions about the influence of psychology on philosophical treatments of the problem of free will. My undergraduate introduction to analytic philosophy's contemporary debate was steeped in compatibilism – Frankfurt, Watson, Davidson, Dennett – with a sceptical spanner in the works being offered by the likes of Honderich and Galen Strawson. But my memory was of a debate carried out almost exclusively in the vocabulary of an abstract belief-desire

psychology typical of the philosophy of mind – abstracted, that is, from too many (messy?) empirical details about human deliberation, decision-making and real-world activity. My sense was that, for a phenomenon whose first-person phenomenology of occurent, conscious, lived volition and action plays such an important role in sustaining the conviction that we have free will, there was a lot being said in the vocabulary of reasons, causes and propositional attitudes without much evident attempt to relate or accommodate to the details of the occurent psychology of real-world agents. How would things look, I wondered, if the data and empirically-inspired theory of psychology and allied disciplines of the mind were brought to bear on philosophical treatments of free will and agency?

In one of life's little accidents[1], I was browsing through some issues of *American Psychologist* that had been discarded by a departing professor when I stumbled across a series of articles in which some psychologists were explicitly addressing themselves to questions about free will[2]. They were, in effect, bringing the data and empirically-inspired theory of Psychology to bear on questions about free will. Three things struck me about these articles[3]. First, they tended to pay little homage to, and were consequently minimally constrained by, the parameters of the traditional philosophical debate. Second, most (if not all) of the contributors seemed to think that psychological science is an explanatory programme that does *not* accommodate free agency – at least, not to the extent that such agency is to be identified with *conscious* human agency[4]. And, third, it was not immediately obvious (at least, not to me) what a philosopher steeped in the traditional debate might say about these views from within psychology, other than perhaps applying labels such as 'incompatibilist' and, perhaps, 'hard determinist', in such a way as to neatly slot the proffered views, opinions and data into a well-organised taxonomy of pre-existing views on free will.

Since this accidental encounter, a steadily growing interdisciplinary literature on free will has come to my attention, including volumes with a more neuroscientific (Libet, Freeman &

[1] I will have more to say about life's little accidents in the penultimate chapter of this thesis.
[2] Issue 7 of Volume 54 (1999) of *American Psychologist*. The articles comprising the special issue or topic were Bargh and Chartrand (1999), Wegner and Wheatley (1999), Gollwitzer (1999), and Kirsch and Lynn (1999)
[3] More specifically, I was struck by these aspects of the contributions by Bargh and Chartrand (1999) and Wegner and Wheatley (1999).
[4] Bargh and Chartrand (1999, p464): "Given one's understandable desire to believe in free will and self-determination, it may be hard to bear that most of daily life is driven by automatic, nonconscious mental processes—but it appears impossible…that conscious control could be up to the job."; Wegner and Wheatley (1999, p480): "…psychological science suggests that all behavior can be attributed to mechanisms that transcend human agency."

Sutherland, 1999) and a more psychological (Baer, Kaufman & Baumeister, 2008) orientation. Some of this literature intersected with and extended my interests and concerns; other parts of it left me unmoved. What crystallised from my engagement with this literature, together with various revisitings of the philosophical debate, was a pair of questions that called for an answer: (i) What place does and should free agency have in Psychology? and (ii) What place do and should psychological views on agency, including free agency, have in a more philosophical debate about free will? These questions provided the immediate impetus for the current project.

A philosophical project intent on somehow answering (or, more modestly, beginning to answer) these questions cannot, however, avoid paying homage to the traditional debate in the way that the contributors to *American Psychologist* saw fit to. Not only would the parameters of the traditional debate need to be outlined in order to better locate and respond to the questions and challenges being raised by the empirically-inspired literature on free agency; the traditional debate itself would need to be revisited in order to consider and evaluate the available positions, and the possibilities for potential (perhaps, optimistically, decisive) movement and progress within that debate.

The cumulative effect of these inspirations, influences and considerations on the shape of this thesis will be the subject of the latter parts of this introductory chapter. Before laying out the plan and rationale of the project in greater detail, however, it is necessary to begin the process of situating the discussion within the parameters of the philosophical debate we have been alluding to above. Having first offered various options for sketching what we could call the 'problem space' of traditional philosophical worries about free will, I will then outline the basic claims of the three dominant positions in that debate[5]: libertarianism, hard determinism and compatibilism. I will then offer a preliminary diagnosis of the state of play in the debate – a stalemate – before laying out the plan for the thesis.

Cosmic Re-runs, Consequences and Sceptical Dilemmas

Philosophical concerns about determinism and free will are perhaps most easily[6] highlighted by way of thought experiments designed to provoke certain intuitions about the nature of

[5] Bearing in mind that finding truly common commitments across different theorists is difficult beyond those at more general, position-defining levels.
[6] If potentially misleadingly…

human agency. The following is one such thought experiment:

Imagine running the 'experiment' that is our universe all over again, right from the (hypothesized) Big Bang[7]. If determinism[8] is true, your life – all your beliefs, desires, emotions, values, projects, experiences, choices, whims, doodles, actions, everything – would turn out exactly as it has done in this world.

This (brief and under-described) 'cosmic re-run' thought experiment is supposed to raise a number of concerns about the implications of determinism for our views about our own agency. I list these in no particular order[9]. First, the experiment suggests an image of oneself – of any human agent – as a very small cog in a very big machine, whose behaviour represents a mere 'unfolding' of a miniscule part of the grand cosmic experiment. This intuition might not only undermine the sense of significance that we tend to attach to our lives and our real (or potential) contribution to the world, but it further portrays that life as a small series of inevitabilities – miniscule turnings of our tiny cog that are just some product of many other turnings of many other cogs, both near and remote in time and space, relentlessly chugging along on the courses that were all set by the relevant initial conditions of the entire system, plus the laws of the universe. This, in turn, easily suggests some further possible intuitive responses to the thought experiment.

The idea that the course of our lives is/was inevitable, all the essential parameters having been laid down and fixed through a combination of the laws of nature and the initial conditions and events in our universe, offends against our sense that it is we – especially in our take on the world and the choices we make in it – who are the owners and originators of a significant part of what we contribute to the course of its history. We feel that such 'originations', these meaningful individual contributions to events that unfold around us, are

[7] Since the focus of the experiment is on the outcome of a single lifetime, starting the re-run at the Big Bang is unnecessary. Any time before the birth of the agent that is the focus of our attention will do.

[8] A lot of ink could be spilt over defining determinism. For my purposes, I am happy to follow van Inwagen (1975) who sees determinism as a conjunction of two claims: "(a) For every instant of time, there is a proposition that expresses the state of the world at that instant; [and] (b) If A and B are any propositions that express the state of the world at some instants, then the conjunction of A with the laws of physics entails B." (van Inwagen, 1975, p186). In a later article, van Inwagen (1989, p400) provides this simpler definition: "Determinism is the thesis that the past and the laws of nature together determine a unique future, that only one future is consistent with the past and the laws of nature."

[9] I also list them without paying any special attention to the myriad qualifications that (especially compatibilist) philosophers might prefer to have inserted from the outset. The point of the thought experiment it scene setting, whether or not the scene thus set involves important omissions and/or 'tricks of light' of various kinds.

in some way crucial to our sense of agency and the ownership of our actions. That such contributions were already 'in the cards' long before we were born does not sit comfortably with our claims of agency and ownership.

That the course of our lives should turn out identically if they were to be 'run' all over again also offends against a different set of experiences and beliefs we tend to hold dear – our sense of the open-endedness that appears to accompany much of our experience, including our experiences of having to make choices and decisions in the face of life's challenges. Far from being predetermined and inevitable, many of our choices have the feel of either being ripe with possibilities for creativity and novelty, or of being weighed down by a sense of what we might call 'agentic vertigo' – a feeling of not knowing how to proceed, or of being torn between multiple potential paths of action, without a clear conviction about which way we should proceed. No matter how 'determined' our efforts[10] in these situations, no matter the degree to which we eventually achieve a level of conviction in these decisions, this apparent open-endedness does not square well with a view of the world in which it was inevitable that things should have turned out thus and so.

Further, this notion of inevitability is difficult to square with our experiences on occasions when our sense of agency is perhaps most strained. On the one hand, when we succumb to weakness of will, we tend to have an experience as of knowing how things should have turned out but for the intervention of some interfering influence. This experience often incorporates a feeling that things would have been different under only marginally (even trivially) different circumstances, as well as the conviction that things really ought to have been different but for a more or less chance happening of some kind, or a lack of sufficient effort on our part. On the other hand, when our will is weak but we succeed in our ends, the sense of *effort* that attaches to these instances of successful action seems at odds with the notion that things *just would have turned out as they did* – indeed, in the context of the experiment, that they would repeatedly turn out the same way – no matter what.

Finally (though much more could, of course, be said), the thought experiment can easily lead us towards worries about notions of responsibility – personal, moral and legal. The picture of a life determined in all its vagaries and complexities long before it has even begun, a life with

[10] In the sense of our determination and effort.

no real alternative branching paths open to it that lead off its singular determined track, suggests a strong *prima facie* case against attaching any meaningful degree of merit or blame to the actions that issue from an agent.

These apparent threats posed by determinism arise without the need to articulate any particular account of free will or agency, although some characteristics of and assumptions about agenthood are at work here, including: the significance of issues of origination, ownership, control, and alternative possibilities; and a basic phenomenology of choice, decision and effort/'perseverance'.

A Classic Argument for Incompatibilism

This cosmic re-run thought experiment, along with the ideas and worries extracted from it, make for a useful if largely unsophisticated and non-technical starting point in coming to grips with the nature, content and parameters of traditional philosophical worries about free will and determinism. Yet the debate itself is a more technical affair, and it is crucially shaped by a range of arguments that have been used to work up one or more of the abovementioned ideas and worries into something more like a proof that determinism and free will are not happy companions[11].

Ideas about events preceding our birth, the laws of nature, the importance of alternatives in choice and action, and the incompatibility of free will and determinism that these might be used to suggest, are perhaps most well developed (within the last 50 years or so of free will debate) in Peter van Inwagen's so-called Consequence Argument[12]. As originally formulated, van Inwagen (1975) asks us to consider a case of a man – a judge, *J* – refraining from a particular action after due rational deliberation and decision. Specifically, we are asked to imagine *J* (at time *T*) refraining from raising his hand where, as a judge in his country, raising his hand would have been effective in waiving the death penalty for a particular criminal. In deciding to not raise his hand, *J* thus decides not to use his powers of clemency, and the criminal is put to death.

[11] Questions about the compatibility of free will and indeterminism will be visited below.
[12] See van Inwagen (1975, 1983).

Since, unlike the 'cosmic re-run' experiment above, van Inwagen asks us to look at a specific case of an individual and their action, it is important to add a few more details about *J* and his decision. For van Inwagen's purposes, we should imagine that *J*:

> …was unbound, uninjured, and free from paralysis; that he decided not to raise his hand at *T* only after a period of calm, rational, and relevant deliberation; that he had not been subjected to any 'pressure' to decide one way or the other about the criminal's death; that he was not under the influence of drugs, hypnosis, or anything of that sort; and finally, that there was no element in his deliberations that would have been of any special interest to a student of abnormal psychology. (van Inwagen, 1975, p191)[13]

On the surface, then, we are presented with a case of a man whose status grants him the power to save another man, who chooses at a given time *T*, and after due consideration in the absence of any duress or compulsion or other failure of agency, not to exercise this power. We are clearly being invited to say of *J* that he *could have raised his hand at T*, even though he did not.

With this background, van Inwagen (1975) then offers what he calls his 'main argument', as follows:

(1) If determinism is true, then the conjunction of P_o and *L* entails *P*.
(2) If *J* had raised his hand at *T*, then *P* would be false.
(3) If (2) is true, then if *J* could have raised his hand at *T*, *J* could have rendered *P* false.
(4) If *J* could have rendered *P* false, and if the conjunction of P_o and *L* entails *P*, then *J* could have rendered the conjunction of P_o and *L* false.
(5) If *J* could have rendered the conjunction of P_o and *L* false, then *J* could have rendered *L* false.
(6) *J* could not have rendered *L* false.
∴ (7) If determinism is true, *J* could not have raised his hand at *T*.

The logical machinery behind the argument is relatively straightforward. Given time T_o, a time some time before the birth of *J*, P_o is the proposition expressing the state of the world at this time T_o. *P*, then, refers to the proposition expressing the state of the world at time *T* – the time at which *J* refrains from raising his hand. Finally, *L* is a proposition containing the conjunction of all the laws of physics.

The critical step in the argument can be stated as a kind of dilemma. If *J* could have rendered *P* false, he could have rendered the conjunction of P_o and *L* false. But then, either *J* could have rendered false a proposition relating to circumstances in the world pre-dating his existence, or he could have rendered a proposition combining the laws of physics false.

[13] That is, van Inwagen (1975) wants *J* to bear all the more obvious hallmarks of compatibilist freedom/s.

Neither option in the dilemma sounds especially plausible, and both sound like they might require a lot more than the ability to raise one's arm.

On van Inwagen's (1975) original version of the argument, he takes it that *J* cannot render P_o false, which leaves *L* as the part of the conjunction of P_o and *L* that *J* could make false if he was to have raised his arm. But in premise 6, he asserts that *J* could not have rendered *L* false. For van Inwagen (1975, p193), it is premise 6 and the connection that it makes between the concepts 'can' and 'law' that lies "at the root of the incompatibility of free will and determinism." The idea behind the premise is simple enough: if any person can render a given proposition false, then that proposition is not a law of physics. Complexities and qualifications aside, the laws of physics just aren't candidates for being rendered false. If a proposition purporting to be a law of physics was rendered false, it would not become a law that turned out to be (or somehow became) false – it would be disqualified from being a law at all. Faced with this impossibility of rendering *L* false, van Inwagen thus asks us to deduce the incompatibility of the truth of determinism and *J's* being able to raise his hand at *T* and, with that, conclude that free will and determinism are incompatible.[14]

In *An Essay on Free Will*, van Inwagen (1983) presents us with a less technical 'overview' version of the argument, this time calling it the Consequence Argument, and stated (in the first person plural) in terms of what is and what is not up to us:

> If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is 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, p16/56)

Here, the explicit emphasis in the argument is less on the conflict between determinism and our having alternative possibilities in action, and more on the idea that our acts should be *up to us* if we are to claim them as free. That is, there is a shift towards questions about the ownership and control of action, in the place of (or, perhaps, just with less emphasis on) issues of alternative possibilities. To claim our acts as free, to claim the kind and level of ownership and control over them that we think compatible with our having free will, requires that these acts are up to us. But the force of the Consequence Argument is to claim that the

[14] Since I am using van Inwagen's argument primarily for expository purposes, I will ignore the variations of the argument presented in van Inwagen (1983). I will also ignore the counterexample developed by McKay and Johnson (1996), acknowledged by van Inwagen (2000) as rendering his original argument invalid. Van Inwagen (2000) offers a repair to the relevant principle that, he claims, saves the argument; he further claims that others, including McKay and Johnson themselves, have made available similar 'repairs'. I will not evaluate these claims here – see van Inwagen (2000) for details.

truth of determinism would make it such that our acts are not up to us – not because we do not feature in the story of these acts at all but because, viewed from the right perspective, they are evidently the consequences of states of affairs that took place before we were born, combined with the laws of nature – neither of which look like things that could possibly be up to us. This conclusion is similar to the concern raised earlier about ownership and origination on the basis of a cosmic re-run; but it is offered to us without the machinery of a thought experiment[15].

The two arguments from van Inwagen presented above capture nicely three central themes running through classical or traditional philosophical debates about free will: themes of ownership, control and alternative possibilities. Each of these themes was, in turn, evident in one or more of the reactions to the cosmic re-run experiment. So far, however, our puzzles and problems about free will have been framed in terms of a tension between determinism and free will. An additional element of the classical or traditional debate involves raising the stakes, or deepening the mystery, by introducing some alleged problems with indeterminism into the mix.

Problems with Indeterminism

One version of this tactic involves framing a sceptical dilemma where, on either horn, free will looks to be in trouble. For example, Peter Lipton offers the following route in to an appreciation of the traditional problem of free will:

> First, everything that happens in the world is either determined or not. Second, if everything is determined, there is no free will. For then every action would be fixed by earlier events, indeed events that took place before the actor was born. Third, if on the other hand not everything is determined, then there is no free will either. For in this case any given action is either determined, which is no good, or undetermined. But if what you do is undetermined then you are not controlling it, so it is not an exercise of free will… [C]onclusion: there is no free will. (Lipton, 2004, p.89)

This sceptical dilemma combines worries about determinism already at work in the cosmic re-run experiment and in the Consequence Argument with a new worry: that if things don't look promising for free will under determinism, then they don't look terribly good either if events in the world are not determined. Specifically, the claim is that if an action of yours is not determined, it is undetermined, and that can't be good news for free will because undetermined acts are not under your (or anyone's) control.

[15] And thus, it is presented without the apparent threat of being accused of priming or manipulating an 'intuition pump' (to borrow Dennett's phrase).

In essence, and leaving technicalities and qualifications aside, Lipton's (2004) sceptical dilemma combines an underlying worry in the Consequence Argument with an underlying worry at work in what van Inwagen (1983, 2000) and others have called the *Mind* argument against libertarianism. This latter argument can be stated as follows:

> If indeterminism is to be relevant to the question whether a given agent has free will, it must be because the acts of that agent cannot be free unless they (or perhaps their immediate causal antecedents) are undetermined. But if an agent's acts are undetermined, then *how* the agent acts on a given occasion is a matter of chance. And if how an agent acts is a matter of chance, then the agent can hardly be said to have free will...[If] an agent is faced with a choice between[, for example,] lying and telling the truth, and if it is a *mere matter of chance* which of these things the agent does, then it cannot be up to the agent which of them he (*sic*) does. (van Inwagen, 2000, p10; italics in original)

The conclusion of the *Mind* argument is stated in terms of things being up to us or not, while Lipton's dilemma states the problem in terms of control; but, at this stage in our discussion, that difference is not all that important. The end result of constructing a sceptical dilemma, or of combining the Consequence Argument about determinism with the *Mind* argument about indeterminism, is a kind of dual or 'hard' incompatibilism: free will appears incompatible with the universe turning out to be deterministic *or* indeterministic. Our claims to having free will are, from this perspective, in a lot of trouble.

Traditional Positions and Responses

Within the confines of the classical or traditional philosophical debate about free will, the positions that emerge in the face of, and in response to, these apparent problems and challenges tend to fall neatly into a relatively clear set of categories. At a fundamental level, these categories are defined by (i) views about the existence or otherwise of free will in human agents, and (ii) views on the potential for conflict between free will and the possible truth of a determinist thesis about the world. Given our introductory foray into the debate, we can also consider (iii) views on the compatibility of free will and indeterminism.

Views under (ii) break down broadly into compatibilist and incompatibilist positions, since (historically) questions about compatibilism have centred on (ii) not (iii). Compatibilists hold that some form of free will is compatible with the truth of determinism, whilst incompatibilists hold that free will and determinism cannot coexist in the same universe. Combining these positions with potential views under (i) gives us the following taxonomy of positions within the traditional free will debate:

- Incompatibilism plus determinism true – 'hard' determinists, who hold that because determinism is true in our world, and because meaningful free will is not possible in a deterministic universe, we do not have free will.
- Incompatibilism plus determinism false – libertarians, who typically hold that we have free will, that free will is not compatible with the truth of determinism, and that the story of how it is that we have free will involves some form of indeterminism in our choices and action.
- Compatibilism and determinism true (or potentially true) – so-called 'soft' determinists or, more commonly, compatibilists, who hold that (at least one or other variety of) free will that is worth having[16] is compatible with the truth of determinism, and we have such free will[17].

Each of these three major positions – 'hard' determinism, libertarianism and compatibilism – will be briefly discussed below in terms of its overall promise and its most evident costs and/or weaknesses. As with the introductory arguments offered earlier, the objective will be to provide a rough-and-ready sketch and assessment of each position, rather than a detailed, nuanced and carefully qualified examination of claims and implications at stake in each case.

'Hard' Determinism

So-called 'hard' determinists, being a variety of incompatibilist, tend to agree with much of what libertarians want to say about the nature of free will, even if only to the point of agreeing that a libertarian conception of free will is *the* conception of free will that is at stake in the debate[18]. They further tend to agree that a free will with the kinds of features desired by libertarians is incompatible with ours being a universe in which determinism holds true. Where the hard determinists and libertarians inevitably part ways is on the issue of the truth of determinism: hard determinists think determinism is true, at least in the only ways that matter to the world of humans and their activities. Thus we have no free will, all appearances and beliefs to the contrary notwithstanding.

[16] To borrow and adapt another phrase of Dennett's.

[17] It is not strictly necessary for a compatibilist to think that we have free will. A compatibilist could argue for a variety of free will that is compatible with the truth of determinism while leaving it open (or even doubtful) as to whether or not human agents qualify as having such free will. But then, if the point of the debate (for those who are not hard determinists) was to save or defend our claims to having free will in the face of the problems we have touched on above, it is not clear what the motivation for such a compatibilist position would be, besides comprising a challenging intellectual exercise.

[18] This need not be the case – a hard determinist might hold that all notions of free will are deeply incoherent, and thus not at all worth saving via either compatibilist or libertarian strategies.

Hard determinism is not a popular position, no doubt in part because it is a deeply sceptical position with potentially far-reaching revisionist consequences for both our self understanding and for a variety of social practices, both micro and macro in scale. Like many sceptical positions, however, it is not easily dismissed, even while the traditional debate is dominated by compatibilist and libertarian voices. In order to avoid a protracted discussion and evaluation of hard determinism, it is thus worth disentangling a number of theses, and dealing with each as economically as possible.

Many of the assumptions made by hard determinists about determinism, causality, and laws of nature are shared with compatibilist positions, and so will not be singled out for criticism here. What marks a position as being one of hard determinism is a commitment to (i) a sceptical or hard incompatibilism about free will, combined with (ii) determinism. The commitment to (ii) is more easily open to criticism than the commitment to (i). Some of our most respected scientific theories (especially those concerning quantum phenomena) tell us that our universe contains real indeterminacy. In that sense, determinism of the global kind articulated in our earlier thought experiment is, according to these theories, literally false. At some level, then, our best scientific theories suggest that hard determinists are wrong to claim that our universe is deterministic.

The hard determinist is likely to respond that determinism is, in some sense, *true enough* of the realm of human affairs for the above criticism to be deflected. Quantum micro-indeterminacies, if there are such things, may well all settle down into sufficiently deterministic macro-patterns at levels of material organisation relevant to human behaviour. Moreover, the hard determinist is likely to hold that it is far from obvious how indeterminacies at the level of microphysics are to have relevance to the realm of human action, let alone to securing free will and responsible agency.

However reasonable this response may seem, it must be countered that the hard determinist's conviction about the truth of determinism, even at levels of phenomena more obviously relevant to human behaviour, far outstrips any evidence that can be offered in defence of such a position. Our state of knowledge, both empirical and theoretical, in psychology, neuroscience, and the interdisciplinary sciences of the mind in general is simply inadequate to justify the claim that real (as opposed to epistemic) indeterminism has no place in the

description and explanation of human activity[19]. At this level, the hard determinist's commitment to determinism has the feel of a promissory note that many feel will not – perhaps cannot – be delivered on[20].

It is the hard determinist's commitment to a sceptical or 'hard' incompatibilism (of the variety highlighted in the sceptical dilemma we encountered earlier) that is more difficult to dismiss: freedom is incompatible with determinism *and* indeterminism. This brand of hard incompatibilism violates not only our intuitions about the nature of human agency, but also the phenomenology of being such an agent. Under such a view, whole swathes of human experience are rendered illusory and misleading in the extreme. Hard determinism thus typically implies a radical revision of our understandings of deliberation, decision, choice and action, and with these the notions of personal, moral and legal responsibility that play a significant role in structuring the social and psychological world we inhabit[21]. And yet, while this means that the stakes are high, and while it would seem *prima facie* that the probability of such widespread and pervasive error and illusion would be correspondingly low, such sceptical incompatibilism cannot be dismissed on these grounds alone.

In the context of introducing the traditional debate and its main positions, however, I will for now set aside sceptical incompatibilism (and, with it, hard determinism) based on the following strategic considerations. The two traditional non-sceptical positions on free will (libertarianism and compatibilism) both try to save the phenomenon of free agency, and our most valued intuitions and beliefs about such agency, in ways that provide much of the bite or impetus for engaging in the traditional debate – that is, the debate is strongly driven by those trying to save the appearance of freedom. Moreover, these non-sceptical accounts are so many and varied that it seems a reasonable (if uncertain) bet that we will be able to save something of the phenomenon, rather than be faced with the prospect of sceptical and radical revisionism[22]. I will, however, briefly return to the threat of sceptical/ hard incompatibilism

[19] See van Inwagen (1983); Mele (2006).
[20] A flaw that it arguably shares with Eliminative Materialism and its sceptical take on folk psychology and folk psychological explanations.
[21] I say 'typically' because it is possible to argue that we can and/or should maintain certain illusions of agency because, for example, human agents and society will function better, or with greater cooperation and justice, etc. if certain illusions of individual freedom and responsibility are sustained.
[22] Again, Mele (2006, p202) is helpful here: "Given the assumption that compatibilism is true, it is very plausible… that human beings sometimes act freely and are morally responsible for some of what they do. Given the assumption that incompatibilism is true, our empirical knowledge is not up to the task of settling the issue whether it is more credible that there are free and morally responsible human agents or that there are no

after we have engaged in greater detail with a sample of these positive positions on free agency.[23]

Libertarianism

The chief virtue of libertarian positions on free will is that they appear to take our lay notions of freedom and agency, and the intuitions provoked by deterministic thought experiments, most seriously. That is, libertarians generally seek to accommodate as much as possible of what we assume to be true of our agency and our notions of freedom into a view of the world that is free from the apparent tyranny of deterministic inevitability. There are many different libertarian positions, and not all of them emphasise/utilise the same features of freedom in action to make their arguments. The apparent virtues listed below are thus not shared by all accounts.

Most libertarians take very seriously our sense of origination for the actions that we paradigmatically regard as free. Humans, they say, are originators of sequences of events that are not predetermined by unfolding patterns and chains of events in the world. The sense given to this notion of origination varies on different accounts (e.g. agent-causation accounts, event-causal accounts, etc), but the essential features tend to be the same. Libertarians object to a view of the human agent (when acting freely) as a mere link in a deterministically unfolding causal chain. Rather, human agents are to be seen as originators of causal sequences not already determined by other causal antecedents. This origination is more than just the ownership of action that tends to be offered on compatibilist accounts – e.g. owning an action as freely chosen because it is consistent with our desires. Libertarian origination is, instead, supposed to capture the sense of spontaneity, creativity, and novelty that accompanies (at least some of) our actions, and give a strong interpretation to our intuition that our freely chosen acts are ones that would not have come about were it not for this creativity, novelty and spontaneity.

A second collection of intuitions that libertarian accounts take seriously are those relating to alternative possibilities in acts that are held to be free. Libertarians take it that our experience of having a range of choices open to us in a decision-making scenario is indicative of real

---

such agents." This suggests something of non-sceptical 'dilemma' about free agency – on one horn, we get compatibilism and, on the other, we get an open empirical question; but, either way, we don't have sufficient reason to opt for sceptical incompatibilism.

[23] See Chapter 5.

potential for choice amongst possibilities; and that deterministic thought experiments like the cosmic re-run, or the circumstances outlined in the Consequence Argument, are incompatible with the reality of alternative possible choices. For the libertarian, re-running certain choice situations would produce, on some occasions, different outcomes despite identical circumstances. So the libertarian is opposed to the idea that determinism could be true whilst it is also true that we have real choice amongst real alternatives. Determinism *a la* a deterministic thought experiment or the Consequence Argument would render the availability of alternative possibilities an illusion. Libertarians will thus seek to secure the reality of alternative possibilities, even at a cost of introducing indeterminism into the process of deliberation and choice (I say 'cost' because of critical comments to follow.)

At a most general level, libertarian positions seem best positioned to do justice to the phenomenology of action and our intuitions about free will by promising a variety of accounts on which (if the accounts are successful) our free choices are real choices (amongst real, accessible alternatives) that are really up to us (both because of novelty/origination and the causal indeterminacy of prior conditions). For this reason alone, such positions are arguably to be preferred to incompatibilist alternatives like hard determinism.

However, the cost of all this accommodation can seem incredibly high, depending on the moves made in attempting to secure freedom of the will. A number of stock objections to various libertarian positions warrant notice here (even if they do not all apply, or apply equally, to all such positions).

First, in pursuit of a more meaty sense of origination and ownership than compatibilists might promise us, libertarians regularly risk replacing one set of puzzles (about freedom) with another set of potentially more mysterious theses. For example, strong positions on origination (such as can be found, for example, within certain agent-causation accounts of free will) raise questions about the interpretability of actions that emerge *ex nihilo*, and about agents who might well be described as unmoved movers. Not only do such agents and their actions create puzzles of psychological interpretation and explanation; but (in the absence of a more general critique of the causal picture of the world that tends to accompany more deterministic approaches) these agents threaten to stand outside the general causal order on which they nevertheless have considerable causal influence.

Such causal puzzles are best highlighted by mentioning one particular tactic used by many libertarians in presenting the case for freedom – namely the introduction of indeterminacy into volitional processes. As many critics have remarked, it is not clear how adding indeterminacy to a causal process like volition can do anything other than *weaken* the link between an agent, their choices and their ensuing actions[24]. Leaving aside questions about determinism, it seems a not unreasonable prerequisite for an action to be the product of a free will that *it first be the product of a will* – that is, in at least one sense of the word 'determined', a free action must first be an action that has been determined according to the will of an agent. And to be determined by the will of an agent, as opposed to issuing from whim or flight of fancy, the action should presumably have appropriate relations to the antecedent psychology of the agent. Adding any amount of indeterminacy into this mixture, whether in the actual process of deliberation and choice (*a la* agent-causation or Searle's recent views) or in preceding psychologically formative events[25] (*a la* Kane), seems to threaten or undermine such links.

Relatedly, certain libertarian accounts may be inclined to suggest an unreasonable degree of autonomy for human agents – specifically, autonomy as regards the influences of culture, personal history, and individual psychology.

A particular concern for many would-be defenders of free will is the amount that is conceded to hard determinism by the libertarian. On one hand, most libertarian accounts appear to imply that it is ultimately the truth or otherwise of determinism that will determine the fate of our freedom. A particularly stark example of this is John Searle's recent take on free will[26], in which he makes it clear that it is up to neuroscience to discover the requisite form(s) of indeterministic brain process that he envisages is (are) required for free will – or else… It is the 'or else' aspect of such a claim that bothers many defenders of free will, especially those of a compatibilist leaning. It is not that science cannot be accorded its due place in telling us how the world really is, perhaps even despite appearances. It is more that the experience and social significance of free agency is such that we would prefer not to have to wait on the

[24] Here and elsewhere in this thesis, I use plural gender-neutral pronouns in the place of singular, gendered pronouns, for generic individuals such as the agent under consideration here. I adopt this usage – I wish it were a widely followed convention – both in order to avoid the clumsiness of his/her and she/he locutions, as well as because the practice is one way in which issues of gender are simply avoided altogether, with only a small grammatical price to be paid.

[25] That are, at least in Kane's case of SFAs, instances of choice and decision in themselves. See Chapter 2.

[26] See, especially, Searle (2001a, 2001b, 2007).

pronouncements of science before we can decide on the sense in and degree to which we are free. Libertarians appear to put great faith in the veracity of our experience of free will; yet they make concessions to determinism (in terms of shared incompatibilist convictions) that seem to concede the possibility that such experiences could be exposed as illusory *tout court*. This raising of the stakes in the debate looks like a high price to pay for securing our *prima facie* intuitions about freedom. Might we not, instead, prefer a safer bet of exploring versions and degrees of freedom that we might lay claim to more or less independently of how the deterministic chips may fall?

Compatibilism

The answer to this rhetorical question provides perhaps the most reassuring rationale for pursuing a compatibilist approach to free will – the idea that, if we succeed, we will have made sense of and secured a place for human free agency no matter what science (or metaphysics) have to tell us about determinism. Although particular compatibilist positions may require some, and perhaps some quite substantial revisions to our ideas about free will, compatibilism *per se* promises to nevertheless secure a meaningful form of freedom even in a deterministic universe.

Moreover, compatibilism promises something that libertarian positions seem to stumble upon time and again – namely to secure a meaningful notion of free will based on an appropriate sense in which the psychology (and other relevant features) of the agent (proximally) determines their actions. Such a determining relationship is not only plausible on compatibilism, but is indeed *required* in order that a position be truly compatible with a globally deterministic causal framework. Thus there is no need and no room for mysterious gaps in the causal process, and no possible puzzles as to the contribution and role of indeterministic processes. In fact, a subtle compatibilism, that allows for the universe turning out to be far less deterministic than our initial arguments and thought experiments assume, may well need to outline the place of indeterminism within the framework of human action, and explain how such indeterminism might constrain but not undermine free agency.

More generally, compatibilism seems to provide a space in which the constraints on human volition and action can be acknowledged and appropriately emphasised without the constant fear that recognising such constraints could expose the illusory status of free will. The complex historical, cultural, social, physiological and psychological determinants and

constraints on behaviour are to be accommodated within an account that nevertheless makes evident the various ways in which humans are *and* are not free. In principle, such a complex and realistic approach to human agency and freedom is to be applauded and encouraged.

Compatibilism represents more of a general commitment, rather than a detailed thesis about determinism, free will, and how precisely the two might fit together. The account of free will (and of determinism) that emerges within any particular compatibilist project will depend on the details of the account produced. In this way, praising and critiquing compatibilist accounts *in general* is a challenging exercise. Having just praised the certain potential virtues of a compatibilist position, it should be noted that not all such accounts will share these virtues to the same degree. Specifically, many compatibilist accounts may eventually, for example, concede too much to determinism (and/or to science) and be too willing to live with a rather anaemic conception of free agency.

Rather than deal with detailed criticisms here[27], I will instead comment on what I take to be some more general problems with compatibilism. An obvious place to start is the intense dissatisfaction that compatibilism has provoked amongst its critics. Kant famously called it a "wretched subterfuge" and "a petty word-jugglery" (Kant, 1788/2010, p96), as in the quote below:

> This is a wretched subterfuge with which some persons still let themselves be put off, and so think they have solved, with a petty word-jugglery, that difficult problem, at the solution of which centuries have laboured in vain, and which can therefore scarcely be found so completely on the surface. (Kant, 1788/2010, p96)

The particular word juggling subterfuge that Kant is referring to here is the idea of settling for a mere "comparative notion of freedom" in the face of the threat posed by determinism:

> According to this, that is sometimes called a free effect, the determining physical cause of which lies within the acting thing itself, e.g., that which a projectile performs when it is in free motion, in which case we use the word freedom, because while it is in flight it is not urged by anything external; or as we call the motion of a clock free motion, because it moves its hands itself, which therefore do not require to be pushed by external force; so although the actions of man are necessarily determined by causes which precede in time, we yet call them free, because these causes are ideas produced by our own faculties, whereby desires are evoked on occasion of circumstances, and hence actions are wrought according to our own pleasure. (Kant, 1788/2010, p96)

One gets the sense that, for Kant, this strategy of substituting real freedom for mere comparative 'freedom' is so woefully inadequate and superficial as an attempt to grapple

[27] As we will see below, I intend to evaluate the prospects for compatibilism in Chapter 3, and develop a serious challenge to it in Chapter 4.

with the problems posed by determinism that he is outraged and deeply offended on behalf of all those who have struggled and laboured in vain to generate proper solutions.

The brand of compatibilism that Kant is most obviously targeting here is the variety that emphasises *negative* freedoms[28] – freedom from various constraints and compulsions. For Kant, such *mere* comparative freedom cannot be adequate to escape the threat to freedom posed by our actions being "a necessary result of the determining causes in preceding time" (Kant, 1788/2010, p96). But it is not just a negative conception of freedom to which Kant is objecting, not merely the absence of external compulsion or constraint that is insufficient to the task at hand. For Kant, it seems, no amount of internal deliberation and judgement, or facts about the internal generation of action, could be up to the task of securing freedom in the face of a backward-extending chain of necessitating determining causes:

> "...it matters not that these are internal; it matters not that they have a psychological and not a mechanical causality...; they are still determining principles of the causality of a being whose existence is determinable in time, and therefore under the necessitation of conditions of past time, which therefore, when the subject has to act, are no longer in his (*sic*) power" (Kant, 1788/2010, p97).

Leaving aside some of the more peculiarly Kantian terminology (and preoccupations), Kant's objection seems simple enough: whatever compatibilism offers, however much it locates and highlights processes internal to the agent, and however much it succeeds in establishing and contrasting a distinctively psychological mode of causation with other merely physical, mechanical modes, yet still what we will have is a chain of causation extending backwards in time that necessitates any given action of the agent, such that the principles of causality at work when the agent acts cannot be said to be in their[29] power. The challenges posed by determinism are not to be so easily avoided[30], otherwise (as Kant reminds us) the puzzles are hardly likely to have survived so long.

Moving on a century or so, James famously rebranded compatibilism as 'soft' determinism, and proclaimed it a "quagmire of evasion". More specifically:

> …we have a *soft* determinism that abhors harsh words, and, repudiating fatality, necessity, and even predetermination, says that its real name is freedom; for freedom is only necessity understood, and bondage to the highest is identical with true freedom…

[28] See Chapter 3 below for a discussion of compatibilist conceptions of negative freedom/s.
[29] As per my earlier footnote, the non-gendered plural pronoun is my preferred method for stating gender-neutral claims, even for single subjects (as in this case). I will take it that, from this point on, any grammatical oddities arising from the mix of single subjects and plural pronouns have been adequately explained.
[30] This is not to suggest that the work of constructing any given compatibilist account is easy. It is the overarching logic of the compatibilist 'move' or 'sidestep' that is being put forward as 'petty word-jugglery'.

> Now, all this is a quagmire of evasion under which the real issue of fact has been entirely smothered. Freedom in all these senses presents simply no problem at all. No matter what the soft determinist mean by it, - whether he (*sic*) mean the acting without external constraint; whether he mean the acting rightly, or whether he mean the acquiescing in the law of the whole, - who cannot answer him that sometimes we are free and sometimes we are not? (James, 1897/2006, p149)

Once again, not unlike Kant, James here appears to be quite fundamentally unsatisfied with the primary compatibilist move of sidestepping questions about determinism. The types or senses of freedom that the compatibilist/ soft determinist would like to tell us about are, for James, simply not problems for us to be troubling with. The compatibilist can have these freedoms, and still the fundamental puzzles and questions about freedom and determinism remain.

Rather than continuing this review of famous critics and their dissatisfaction with compatibilism, let us instead note that a similar concession to compatibilist notions of freedom was made in van Inwagen's (1975) setting up of his 'main' argument for incompatibilism. Recall that, in the case of our judge *J*, we were to allow that:

> …was unbound, uninjured, and free from paralysis; that he decided not to raise his hand at *T* only after a period of calm, rational, and relevant deliberation; that he had not been subjected to any 'pressure' to decide one way or the other about the criminal's death; that he was not under the influence of drugs, hypnosis, or anything of that sort; and finally, that there was no element in his deliberations that would have been of any special interest to a student of abnormal psychology. (van Inwagen, 1975, p191)

What is this except a long and, one assumes, non-contentious list of external and internal conditions for agency – freedoms of a sort, as James seems willing to call them – that all sides to the debate can agree must be in place before we can begin to ponder whether or not, in the face of determinism, agents like *J* can be said to be acting freely?

This kind of deep and fundamental dissatisfaction with compatibilism is the obvious converse to the dissatisfaction that non-libertarians feel about the prospects for securing free agency by somehow adding indeterminism into moments of volition. But the libertarian critic of compatibilism is likely to point out that at least they (the libertarians) are taking seriously the problem posed by determinism, and are grappling with the resulting puzzles and challenges that arise from trying to break the hold of the backward-extending chain of causality highlighted by Kant, van Inwagen, Lipton, and so many others. From this perspective, the compatibilist response represents a failure to acknowledge that there is a problem, even if everyone else thinks there is.

The critics' dissatisfaction is not limited to the compatibilist refusal to see significant problems arising out of the backward-extending chain of necessitating causation. Non-compatibilists tend to be equally discontent with the compatibilist attitude towards alternative possibilities where, at best, these are interpreted along conditional lines such that (again, as far as critics are concerned) the puzzles arising from determinism are not given their full force or, at worst, alternative possibilities are specifically *denied* any special significance[31]. For most non-compatibilist believers in free will, having alternative possibilities means having real, accessible, open, branching paths for choice and action extending into the future. On determinism, that future is unique, fixed and unalterable without going back to the initial conditions of the universe, or altering the laws of physics. As James put it, in typically evocative terms, determinism:

> …professes that those parts of the universe already laid down absolutely decree what the other parts shall be. The future has no ambiguous possibilities hidden within its womb: the part we call the present is compatible with only one totality. Any other future complement than the one fixed from eternity is impossible. (James, 1897/2006, p150)

As James puts it a little later in his lecture, on determinism there is necessity and impossibility and nothing else. For the critics, this is the appropriate perspective from which to think about alternative possibilities in relation to determinism and, therefore, this *is* the problem of free will, at least viewed through the lens of issues of alternative possibilities. Such critics find the compatibilist suggestion that conditional readings of alternative possibilities can make this problem go away deeply unsatisfactory[32].

The State of Play: Stalemate

The above should serve adequately as an overview of the basic positions in the classical or traditional debate. It is, as far as philosophical debates go, a relatively highly charged affair where the stakes are high: these are not merely questions of metaphysics and philosophy of mind, but questions about, and with potential implications for, ethics, law, politics, education, and, most generally, our conceptions of ourselves as selves and agents.

[31] Compatibilist views on alternative possibilities, including Frankfurtian and Dennettian attempts to deny their significance, will be discussed in detail in Chapter 3.

[32] Neither the critic nor anyone else need find no value in what compatibilists have to say in their conditional accounts. The critic is just not convinced that anything could be said along these lines that will take proper cognisance of the original problem.

What, then, can we say about the state of play in the traditional debate in advance of exploring the questions posed earlier about the interplay between free agency and psychology[33]? Writing in 1986, Thomas Nagel admitted to changing his mind every time he thought about the problem of free will, and asserted that nothing that might resemble a solution to the problem had yet been proposed[34]. This combination of a philosopher's confession with a blunt and pessimistic assessment of the state of play tells us, I think, two important things about the traditional debate about free will. First, the problem (and the resulting debate) seems to involve a deep cognitive tension that, for many[35], has something like the effect of the transition between different gestalts in an ambiguous figure. Second, it is easily possible to feel and pronounce profound dissatisfaction with *all* the apparent options and extant positions in the debate without this dissatisfaction being motivated by excessive scepticism[36]. With the exception of a small minority of hard determinists, everyone agrees that we have free will, but many just don't know what they can sensibly and consistently say about it.

The deep cognitive tension just mentioned is well illustrated in form, if not necessarily in content, by Kant's presentation of the Third Antinomy of Pure Reason in the *Critique of Pure Reason*. The thesis of freedom is presented (physically) alongside the antithesis that would deny the existence of freedom and only allow causation in accordance with the laws of nature, as are the proofs offered for each, and the comments that follow[37]. It is as much in this physical presentation of the antinomy as in anything else that Kant says about it (and the other antinomies of pure reason) that we are being encouraged to acknowledge the pull of each – the seeming inevitability of each – in direct contradistinction to each other. We must somehow be committed to thesis and antithesis at the same time, and yet we cannot see how this could be.

---

[33] That is, the questions: (i) What place does and should free agency have in Psychology? and (ii) What place do and should psychological views on agency, including free agency, have in a more philosophical debate about free will?

[34] I quote Nagel (1986) directly, and more extensively, on these issues in Chapter 4.

[35] I say 'for many' because there are those committed participants to the debate for whom the tension I describe has been somehow resolved – they no longer feel or acknowledge the pull of the alternative perspective (gestalt) one can take on the problem.

[36] Perhaps consciousness represents one rival to free will for generating philosophical pessimism in the absence of scepticism. Here, for example, is Fodor's assessment of the state of consciousness studies in the early 90s: "Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be like to have the slightest idea about how anything material could be conscious. So much for the philosophy of consciousness" (Fodor, 1992, p5).

[37] I suspect the presentation varies in different editions, but my understanding is that the presentation of thesis and antithesis in parallel columns of texts is Kant's intended format of presentation. See, for example, pages 405-411 of Kant (1781/1787/2007).

Nagel's confession to changing his mind every time he dwells on the problem of free will can equally be understood as a tension between two compelling, indeed apparently necessary perspectives, as in my gesture towards shifts in the perception of ambiguous figures. The problem of free will in the face of determinism is arguably best appreciated from a global (even universal) perspective: cosmic re-runs, chains of causation extending back to before one's birth (or to the big bang), a universe with one and only one path into the future. An obvious antithesis here is one best appreciated at a much more local, individual level. At the level of our experience of own agency, especially in the space of choice and impending decision, it seems obvious, compelling, beyond question that things *are* up to us, and ripe with possibilities, in ways that have somehow disappeared from view at the global perspective – much as the vase disappears as we switch perceptual gestalts to the two faces in the Rubin vase/face figure. As a Kantian might put it, as agents we cannot but act under the idea of freedom[38].

As should be clear at this point in our discussion, this deep tension, framed in this manner, motivates different responses amongst participants in the debate. Libertarians find the antithesis compelling and so, in so far as they view the truth of the thesis as incompatible with the antithesis, they reject the thesis and take on the burden of making sense of free will while rejecting determinism. Compatibilists also find the antithesis compelling, but seek to account for its truth in ways that resolve any apparent conflict with the thesis.

At the same time, compatibilists and hard determinists (or sceptical incompatibilists) reject the libertarian move by pointing to a different kind of deep cognitive tension at play in the debate: the tension highlighted in the sceptical dilemma discussed earlier (Lipton, 2004). From this perspective, the libertarian is claimed to have nowhere to go since the only alternative is indeterminism, and the undetermined acts of an indeterministic form of agency are rejected as candidates for securing an adequate sense in which things could be up to us.

The result of these deep cognitive tensions – these deep metaphysical tensions – is, I think, a stalemate. It is a stalemate that is not necessarily reflected (either historically or contemporarily) in the numbers of thinkers who might claim allegiance to one or other of the

[38] See, for example, Allison (1990).

main positions in the debate – in the absence of detailed survey data, it seems fair to say that hard determinism is represented by a very small minority, while compatibilism probably enjoys a comfortable majority (especially in the modern era), perhaps almost to the point of orthodoxy in late 20th/ early 21st century Anglo-American philosophy[39]. But, certainly, it is a stalemate in terms of any decisive, widely accepted move to promote or undermine either of the two primary positive positions (compatibilism and libertarianism). It is this stalemate that, arguably, both underpins and is so well captured by Nagel's frank confession.

Prospects for Movement, Prospects for Change

Diagnosing a stalemate in the traditional debate about free will is, on the surface, a pessimistic if unsurprising evaluation based on the considerations and arguments touched on thus far. But pessimistic evaluations, in the context of this brief introduction and assessment of the debate, are useful to highlight because they help provide a dual motivation for the current project: (a) they motivate a revisiting of the contemporary debate in order to evaluate the prospects for movement and progress, given developments over the last ten to twenty years; and (b) depending on the outcome of (a), they may motivate a search for fresh perspectives, data and arguments that could provide movement, the promise of progress, and/or (at the very least) directions for research on free agency that have tended to be neglected within the parameters and constraints of tradition. In broad terms, it is the challenges pointed to in (a) and (b) that this project intends to take up.

In Part I, I take up the task pointed to in (a) – engaging with certain significant parts of the traditional debate over the last two decades, looking for signs of movement beyond the stalemate sketched above. I begin (in Chapter 2) by considering the work of Robert Kane[40], whose attempts to develop an event-causal brand of libertarianism have done much to revive interest in positive/ constructive incompatibilist accounts of free will – perhaps most notably because his event-causal account, his occasionalism with regard to the critical events that he thinks might secure an agent's 'Ultimate Responsibility', and the general clarity and transparency of his account, have done much to dispel the air of mystery so often encountered surrounding the crucial parts of agent-causal libertarian accounts. I defend Kane against the critique of his work developed in Dennett's (2003) *Freedom Evolves,* in part because I think

[39] See Nichols (2007) for an interesting take on the rise of compatibilism, and a quantitative approach to the history of philosophy.
[40] See especially Kane (1996, 2002a).

that Kane's account cannot be easily dismissed on the charge of 'chance is just chance' that compatibilist thinkers (in particular) are so fond of employing against libertarians[41]. However, in my own critical assessment of Kane, I argue that he has not done enough to persuade us of undetermined Self-forming Actions (SFAs), and conclude that the compatibilist suspicion about the problems that arise from inserting indeterminism into critical moments of volition appears largely correct.

In Chapter 3, I turn to the difficult task of evaluating the prospects for compatibilism – difficult because, unlike the relatively clear division between agent- and event-causal accounts under libertarianism, it is much more difficult to propose a neat taxonomy of compatibilist accounts that would help simplify both the presentation and evaluation of contemporary compatibilist thinking[42]. Nevertheless, I argue that there is much good sense to be found in aspects of compatibilist thinking about agency and volition, especially as long as one is focussed on the more local level of how choices and actions relate to and flow from the character and psychology of the agent. I conclude the chapter by noting a number of lingering questions and concerns about which non-compatibilists might remain uneasy.

The real work of evaluating the prospects for compatibilism, however, is done in Chapter 4, where I consider, develop and defend an argument (due to Mele, 2006) – the zygote argument – against compatibilism as an adequate account of free agency. As already flagged in our introductory discussions of compatibilism, determinism and free will, the real threat posed by determinism is best seen at a global level. In essence, the zygote argument is an attempt to expose or reframe the difficulties posed by global determinism, about which a committed compatibilist is evidently unconcerned, through the addition of *manipulation* of an agent that (if the argument is successful) respects notions of compatibilist freedom and autonomy. I defend the zygote argument against a series of compatibilist objections and replies, and conclude that compatibilism fails because it allows the possibility of agent manipulation despite an agent being able to lay claim to compatibilist freedom. The incompatibilist suspicion that one cannot secure freedom in a universe where determinism reigns also, then, turns out to be correct. This concludes Part I of the thesis.

---

[41] That is, the ideas about chance, randomness and indeterminism typically at work in the thinking behind the *Mind* argument discussed above.

[42] For example, as we will see in what follows, one can distinguish 'structural' and 'genetic' compatibilist accounts, but the differences in their synchronic and diachronic emphases need not translate into differences in, say, their views on alternative possibilities.

I begin Part II by revisiting the diagnosis of the traditional debate developed in this introduction, and argue that the stalemate described at the outset looks more like a chronic and unproductive impasse in the light of the failures of Kane's event-causal libertarianism, as well as the autonomy-focussed brand of compatibilism targeted in the zygote argument[43]. There are insights to be had on either side regarding the nature of human agency, but those insights are enslaved to and constrained by the problem space of the traditional debate.

Thus, in Chapter 5, I propose setting aside the framework of the traditional debate in order to consider questions about free agency through an alternative set of lenses. There are issues raised by both sides of the traditional debate that deserve our attention, and the state of impasse and stalemate makes the prospect of successfully addressing these concerns seem dim. My proposal, adverted to earlier in this introduction, is that we look to the psychological literature to help both in reframing the key issues in the debate, as well as in generating movement and progress towards a more empirically-informed, but also philosophically and theoretically defensible account of free agency. The empirical literature stemming from psychology and allied sciences of the mind may help us illuminate some of the key concerns about free agency that are at stake in the traditional debate while also pointing to avenues we could explore to address these concerns, even if we do so without resolving the traditional debate itself.

The alternative perspective I propose involves defending our free agency against the claim that we might be some kind of *Agent Automatons* whilst at the same time resisting the threat (or temptation) of securing our claims to freedom by presenting ourselves as variants of *Hyper-reflective, Hyper-rational Agents*. We do not need a thesis of global determinism in order to pose and appreciate threats to our claims of free agency – empirically-inspired claims about automaticity, the timing of conscious intention, and puzzles and illusions of conscious willing are both local *and* sufficient for us to appreciate potential threats to our claims of agency, independent of any convictions about determinism/ indeterminism and compatibilism/ incompatibilism. That is, some have turned to the empirical sciences to suggest that we may be much more like *Agent Automatons* than is consistent with an idea of

[43] While the zygote argument, as we will see, does focus strongly on autonomy-based compatibilist accounts, there is good reason to think that it can generalize to less-demanding compatibilist accounts, especially structural/ hierarchical accounts.

ourselves as *free agents*, and these threats require a response irrespective of one's libertarian or compatibilist leanings[44]. And yet, since the challenge posed more or less cuts across the lines of the traditional debate, it is not clear that the traditional camps within that debate are well positioned to respond to the challenge. We need to lay claim to our status as free agents with something more than, and something more empirically nuanced than, moments of agent-causation, SFAs, or reason-sensitive agency. This is the first context in which Psychology and allied empirical sciences of the mind promise fresh perspective on questions about free will.

At the same time, certain philosophically- and empirically-inspired insights into the nature of real-world conscious, reflective, deliberative agency jointly warn against grounding our response in claims about reflective rational agency that cannot be sustained in the light of the facts. If we are to wrest claims of ownership and control of our lives as agents back from the sceptics who would portray us as *Agent Automatons*, we dare not risk doing so by suggesting that we are, or that we aspire to be, *Hyper-reflective, Hyper-rational Agents*. This is a second context in which the empirical study of the mind and agency provides fresh perspective on both the nature of, and constraints on, our agency. What we need are directions for exploring and articulating the form/s that our real-world version of conscious self-governance might take while steering clear of these extremes. Here we find a third context in which Psychology and its allied disciplines promise not only fresh perspective but also positive avenues for constructing accounts of free agency that move away from classic libertarian and compatibilist frameworks and desiderata.

In pursuing these goals in Part II, I begin (in Chapter 6) by laying out the alleged threats to free agency posed by empirical work on automaticity, the timing of conscious intention, and illusions of conscious willing. While some share my conviction that psychology and psychological research might allow us to make progress in our attempts to make sense of agency and freedom, there are others within and outside of psychology who think some of its results make free agency seem more unlikely than abstract arguments about freedom and determinism. The experiments and data discussed in Chapter 6 have been used to pose just such sceptical arguments about our agency. The need to recognise and respond to these data

[44] I take it that hard determinists are unlikely to be too concerned about the claim that we are agent automatons – this just adds more (empirical) grist to their mill.

and arguments forms a significant part of my case for developing a more psychologically-informed account of agency.

In Chapter 7, I begin responding to these empirically-based sceptical arguments with a few significant correctives and clarifications that begin to diffuse at least some of the apparent threat to our agency. However, the central questions raised by the sceptical arguments of Chapter 6 involve issues of *control*, and these cannot be adequately addressed or responded to without considering in some depth the varieties and forms of control that we might find, as a matter of empirical fact, in complex biological systems like human bodies. This task is taken up at length in Chapter 8.

While Chapter 8 argues that we should expect to find multiple forms and systems of control at work in human agents, including a variety of decentralised and distributed systems of control, we nevertheless do make claims as to the significance of distinctively human forms of control – specifically, conscious control – over our lives, and these claims are critical to our free agency. In pursuit of a realistic and empirically-informed account of such self-governance might look like, Chapter 9 begins this task by (perhaps paradoxically) sounding a number of cautions about avenues we should avoid lest we fall for the trap of claiming and/or aspiring to be hyper-rational, hyper-reflective agents.

With these cautions in mind, Chapter 10 takes up the positive challenge of exploring less-travelled avenues down which we might seek to locate the distinctively human forms of generation, ownership and control of action that secure and justify our claims of free agency. Specifically, I examine *imagination* and various *externalised* aspects of mind as promising spaces in which to explore, locate and defend significant degrees of freedom in human agency. I also bring the discussion full-circle by taking up a challenge raised in Chapter 5 by those who remain committed to the framework of the traditional debate – namely, the question of where (given the incompatibilism implied by my arguments of Chapter 4) I think indeterminism might fit into an account of human free agency, given my rejection of traditional libertarian attempts to insert it into moments of volition[45].

[45] More precisely, I expand on and defend my speculative proposal in Chapter 10, having outlined it in response to the challenge when it was first raised in Chapter 5.

Finally, Chapter 11 summarises the conclusions and contributions of each chapter, of Parts I and II as significant portions of the whole, and of the thesis as a whole, before (re-)advertising what I take to be the most promising avenues for further work stemming from the arguments and claims of the current project.

Chapter 2

A New Voice for Indeterminism: Kane's Account of Free Will

The introductory discussions of the previous chapter suggest that the topography of the free will debate within philosophy is a relatively well traversed and mapped-out landscape, populated by and organised into a stalemate of rival factions without much obvious prospect for movement. Yet the diagnosis of a stalemate cannot be made or maintained simply on historical grounds using well-worn arguments that do not necessarily address themselves to the details of particular extant accounts. It is to the task of a more detailed examination that we must, therefore, now turn.

As I indicated in the Introduction, Robert Kane's work in developing a strong and clear event-causal libertarian account of free will has arguably done much to revive interest in positive incompatibilist accounts of free agency. His event-causal approach is less mysterious than the agent-causal accounts offered by many of his libertarian counterparts[46]; and his occasionalism regarding the frequency of his responsibility-grounding self-forming actions makes it such that there is less indeterminism to be found in his overall picture of the free human agent[47]. Evaluating his account therefore presents perhaps our best opportunity to judge the prospects for progress within traditional libertarianism.

I begin the chapter by outlining the important details of Kane's account. I then present Daniel Dennett's extended critique of Kane in his 2003 book *Freedom Evolves*, and defend Kane against the bulk of Dennett's criticisms. Finally, I develop my own argument against Kane's account, specifically claiming that his indeterministic moments of causal regress-stopping within self-forming actions (SFAs) cannot do the work he needs them to do in distinguishing genuine SFAs from non-SFAs. Despite a valiant effort, his attempt to secure a libertarian account of free will by inserting indeterminism into these special moments of self-forming choice fails.

---

[46] See, for example, Clarke (2003a, 2003b) and O'Connor (1995a, 2000).
[47] The significance of this occasionalism is discussed in due course below.

Ultimacy, Ownership, and Self-Forming Actions

The emphasis in Kane's theory[48] is, from the outset, on notions of creativity, origination, ownership, and associated notions of responsibility. We can see this clearly in his preferred definition of free will:

> [Free will] I define as "the power to be the ultimate creator and sustainer of one's own ends and purposes." (Kane, 2002a, p.223)

It is also worth noting the emphasis that has been placed on ends and purposes. While ends and purposes must clearly be manifest in choice and action, Kane's initial emphasis is on the creation and sustaining of ends and purposes. As we will see, this shift of emphasis from choice and action to responsibility for ends and purposes will do a fair amount of work within the theory.

This shift in emphasis is again evident in the way that Kane sets up his account in contrast to much that has been written in the incompatibilist tradition. In particular, Kane holds that the condition of alternative possibilities (AP) typically emphasised by incompatibilists, and the source of some of the most long-standing arguments between incompatibilists and their compatibilist opponents, is insufficient to make a case for the incompatibility of free will with determinism. Instead, Kane offers us his condition of ultimate responsibility (UR) as the crucial principle at stake in the argument between compatibilism and incompatibilism:

> The basic idea [behind the condition of ultimate responsibility or UR] is this: to be ultimately responsible for an action, an agent must be responsible for anything that is a sufficient reason (condition, cause or motive) for the action's occurring. If, for example, a choice issues from, and can be sufficiently explained by, an agent's character and motives (together with background conditions), then to be *ultimately* responsible for the choice, the agent must be at least in part responsible by virtue of choices or actions voluntarily performed in the past for having the character and motives he or she now has. (Kane, 2002a, p.224)

We can immediately see, in the formulation of UR, that Kane is making room for the possibility that any given choice may be sufficiently explained – and, in this sense, determined – by an agent's character and motives together with various relevant background conditions. That is, his indeterminist theory appears at the outset to have shifted focus from any and all choices (where we might ask in every case about the availability of alternative possibilities) to some subset of "choices and actions voluntarily performed in the past" that

[48] I use Kane (2002a) as my primary source because of both its clarity of presentation, as well as because it reflects a version of his account that has been shaped by responses to Kane (1996).

played some significant role in establishing an agent's character and motives, and (if Kane's account is to succeed) for which we can attach a special kind of responsibility[49].

Kane thus thinks that UR should be the focal point of the argument between compatibilists and incompatibilists. But compatibilists are clearly in the business of providing accounts of how responsibility of the right kind can be attached to choices and actions in the right way so that agents become responsible for acts of their own free will in a context where determinism holds sway. So Kane needs to say something more about UR, and how it functions within his theory to produce a substantive clash between determinism and free will – an argument that traces the source of incompatibility to something other than a general problem about the availability of alternative possibilities. The core of the argument for incompatibility becomes something like the following:

> If agents must be responsible to some degree for anything that is a sufficient cause or motive for their actions, an impossible infinite regress of past actions would be required unless some actions in the agent's life history (SFAs [Self Forming Actions/ Self Forming Choices]) did not have either sufficient causes or motives (and hence were undetermined). (Kane, 2002a, p.225)

As Kane acknowledges, this line of argument in turn raises serious difficulties about, for example, how we can talk of free and responsible actions when these actions are said to lack, in some significant sense, either sufficient causes or motives. He further thinks that incompatibilists are charged with the task of explaining how such actions can be accommodated in the natural order of things. The initial claim, however, is simply that this is where the significant work must be done, rather than perpetuating a stalemate in the debate about alternative possibilities.

The focal issue has thus now shifted to one of how to get the buck to stop with the agent without conceding that all such character-forming choices and actions are *determined* by who the agent is and what they do, and yet despite this lack of determination of character formation, the whole thing is nevertheless still 'up to us' in some significant sense. The issues involved comprise what Kane calls the 'intelligibility question': if free will is not compatible with determinism, how is it any more compatible with indeterminism? As noted in my introductory survey of the terrain of the traditional free will debate, a central problem for indeterminism has always been the problem of explaining how the addition of indeterminism

[49] By limiting the need to posit special moments of indeterminism to a subset of choice situations, Kane is favouring what I have been calling an 'occasionalist' brand of libertarianism; whereas agent-causal libertarians, and Searle's (2001a, 2001b, 2007) recent advocacy for an indeterministic account of freedom, tend to see indeterminacy as a regular feature of choice situations.

into the mix of character, motivation, deliberation and choice can make for action that is more, and more freely, the product of the agent's will. Kane has a rather low, and candidly expressed, opinion of incompatibilist or libertarian attempts to address what he calls the intelligibility question.

> Realizing that free will cannot merely be indeterminism or chance, they [libertarians] have appealed to various obscure or mysterious forms of agency or causation to make up the difference… Many…respectable philosophers continue to believe that only some sort of appeal to mind/ body dualism can make sense of free will…[And] the most popular appeal among philosophers today is to a special kind of *agent- or immanent causation* that cannot be explained in terms of the ordinary modes of causation in terms of events familiar to science… I call these familiar libertarian strategies for making sense of free will "extra factor" strategies. (Kane, 2002a, pp.226-227, italics in original)

Kane thinks that libertarians would be on much firmer ground if their theories were constrained, from the point of view of ontology, to exactly the same kinds of entities, states and processes that need to be posited in compatibilist accounts. Rather than postulating causal factors and modes of causation that are not to be found elsewhere in science, Kane is going to make a thorough go of producing an indeterminist account that draws, as far as possible, on what we already know about the world described by science. In this sense, at least, it will be a *naturalistic* theory of free will that is nevertheless libertarian in its outlook and conclusions.

As already intimated above, one of Kane's most significant departures from the indeterminist tradition – one that allows him a special concession to compatibilists – is that he does not require indeterminism for every action, or in every instance of choice.

> …[I]ndeterminism does not have to be involved in all acts done "of our own free wills" for which we are ultimately responsible… Not all such acts have to be undetermined, but only those by which we made ourselves into the kinds of persons we are, namely "self-forming actions" or SFAs…[T]hese undetermined self-forming actions or SFAs occur at those difficult times of life when we are torn between competing visions of what we should do or become. (Kane, 2002a, pp.227-8)

Identifying self-forming actions and choices as only a subset of (potentially infrequent, if not rare) events in an agent's life in which indeterminism needs to play a role has a number of important implications. First, while Kane is, by his own admission, as committed as any other indeterminist to it turning out (empirically) to be the case the relevant kind of indeterminism required by his theory is in fact to be found in the world (specifically, in the functioning of the brain and body), his theory will require considerably less indeterminism than is required by most other libertarian or indeterminist theories. Agent-causal theorists who postulate a special kind of agent-causation in most if not all instances of human choice and action will require just that many 'gaps' in the explanatory stories that can be told in terms of the

'ordinary modes of causation' encountered elsewhere in science. John Searle, in his recent indeterminist forays into the free will debate[50], acknowledges that his account requires neuroscience to find the relevant and frequent indeterministic gaps in neural functioning that will make space for, or map onto, the numerous gaps that his phenomenologically-driven account seems to point to. Kane, by contrast, requires something of a smattering of genuinely undetermined life choices for which we need to find real indeterminacy in the world.

A second related implication is that Kane is not offering an account of free will that attaches indeterminism to the phenomenology of deliberation, choice and action *in general*. The account does not, it seems, offer anything obvious that we might use to *explain*, for example, the 'gappiness' that Searle sees as the central explananda in his account of rational agency and free will[51]. Kane will talk about deliberation, conflicting motivation, and effort; but while these may be of quite general relevance to human volition and action, Kane's theory only locates a significant role for indeterminism in SFAs.

A third implication, anticipated in my comments on Kane's principle UR, is that Kane can explicitly allow for important life choices being *determined* by an agent's character and motivational set without running into the kinds of problems that other libertarians might encounter. Kane specifically discusses Dennett's example of Martin Luther proclaiming "Here I stand. I can do no other" when he refused to recant his heretical views. Libertarians who see indeterminism everywhere in human choice, and/ or whose theories require alternative possibilities for any given action if it is to qualify as free, are justifiably uncomfortable when confronted with this kind of case. But the case presents no problem for Kane – at least as long as he can say something about undetermined SFAs in Luther's life history that can ground an attribution of ultimate responsibility. More generally, Kane's account allows that 'snap' decisions and choices made without deliberation can sensibly qualify as acts of the agent's free will, again on condition that such choices issue from a character and motivational set that is appropriately grounded in undetermined SFAs.

As is becoming clear, it is these self-forming actions and choices that form the centrepiece of Kane's theory. We need to be very clear on exactly what Kane thinks is going on when an

[50] See Searle (2001a, 2001b, 2007).
[51] *ibid.*

agent finds themselves in one of these special kinds of choice situations. His description is as follows:

> There is tension and uncertainty in our minds about what to do at such times, I suggest, that is reflected in appropriate regions of our brains by movement away from thermodynamic equilibrium – in short, a kind of "stirring up of chaos" in the brain that makes it sensitive to microindeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation is thus reflected in the indeterminacy of our neural processes themselves. What is experienced internally as uncertainty then corresponds physically to the opening up of a window of opportunity that temporarily screens off complete determination by influences of the past. (Kane, 2002a, p.228)

We are asked, then, to think of choice situations in which we face a conflict of motivation, characterised by tension and uncertainty about which direction or path we should take. Kane's proposal is that this tension, conflict and uncertainty is, in some way, echoed at a neural level in a "stirring up of chaos" that makes the brain sensitive to microindeterminacies – i.e. quantum indeterminacies – somehow amplified in the non-linear dynamics of the functioning brain. We will return to the role played by quantum indeterminacy and the mathematics of chaos in due course. For the moment, we need to get more clarity on the kind of choice situation an agent might be facing when confronted with one of these special opportunities for self-forming choice and action.

Kane's favoured example is that of a businesswoman making her way to work where she has an important meeting to attend. On her way, she witnesses an assault taking place in an alley leading off the street she is walking along. Kane asks us to imagine her conscience or moral convictions inclining her to stop and, at least, call for help; whereas her career ambitions and commitments to her work tell her that she cannot miss the meeting. If she stops, she will be late for the meeting; if she waits till she gets to her office to make a phone call for help, she is unlikely to be of much assistance to the assault victim. Her options, in terms of action, are mutually incompatible; her motivations for either course of action, Kane suggests, are based on different and incommensurable reasons. Kane himself describes the example in a morally-laden manner, setting up a struggle between her conscience's demand that she stop and help and the "temptation" to go on to her meeting. But I take it that the more important aspect of the case, from the point of view of setting up an example of self-forming choice, is the incompatibility and incommensurability just described, and the associated internal conflict experienced by the businesswoman.

In a significant sense, the central task for Kane is to make the argument that the indeterminacy involved in SFAs – indeterminacy arising out of a conflict of motives – can issue in choices that are nevertheless willed. In other words, Kane needs to make the argument that the choice or decision involved in an SFA which, by his own lights, is undetermined, can nevertheless be an expression of the will of the agent, and thus an action that is properly considered rational and voluntary. Kane wants to make the following claim plausible:

> When we… decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we *make* one set of competing reasons or motives prevail over others then and there *by deciding*. (Kane, 2002a, p.228)

But just how can we make sense of the businesswoman 'making' one set of reasons prevail, and 'deciding' what to do, when there is indeterminism involved? This kind of question has always been something of an Achilles heel for indeterministic accounts of free will.

Kane has a number of examples that he uses to make the required argument. First, he uses two examples – that of an assassin and of an angry husband – to try make plausible the claim that intentional doing and responsibility can 'survive' the intervention of indeterministic processes in the performance of an action. Here is his description of the case of the assassin:

> Consider an assassin who is trying to shoot the prime minister, but might miss because of some undetermined events in his nervous system that may lead to a jerking or wavering of his arm. If the assassin does succeed in hitting his target, despite the indeterminism, can he be held responsible? The answer is clearly yes because he intentionally and voluntarily succeeded in doing what he was *trying* to do – kill the prime minister. Yet his action, killing the prime minister, was undetermined. (Kane, 2002a, p.229)

As an alternative example, Kane asks us to consider the example of a husband who flies into a rage while arguing with his wife. He swings his arm down onto his wife's favourite glass-topped table, intending to break it. Kane asks us to consider the following possibility:

> Again, we suppose that some indeterminism in his outgoing neural pathways makes the momentum of his arm indeterminate so that it is undetermined whether the table will break right up to the moment when it is struck. Whether the husband breaks the table or not is undetermined and yet he is clearly responsible if he does break it. (Kane, 2002a, p.229)

We have, then, two candidate examples of actions whose success in achieving their intentional outcomes is, according to Kane, undetermined because of the hypothesised indeterminacy that arises in the relevant (motor) neural processes; and yet we would hold each agent responsible for their actions should they succeed (and, certainly in the case of the assassin, even if they didn't). Kane goes further in considering either agent offering an excuse like the following: "I'm not responsible. I didn't do it – chance did!" Kane invites us to agree with his judgement that we would reject this excuse out of hand. Both agents succeeded in

doing what they were voluntarily intending to do. While the indeterminism in their neural pathways made the outcome of their efforts undetermined – the indeterminism constituting a potential obstacle or hindrance to their success – their successful actions achieved their intended objective, and we would hold them responsible accordingly.

Kane has a further example in mind – one involving 'internal' activity rather than externally directed action. He asks us to imagine ourselves trying to solve a mathematical problem. Again, he postulates some degree of indeterminacy in our neural processes that forms "a kind of chaotic background." (Kane, 2002a, p.229) So, whether we solve the problem or not is, by his lights, undetermined because of the potential interference of the postulated "distracting neural noise." Kane asks us to share his interpretation of this cognitive scenario:

> …if you concentrate and solve the problem none the less, we have reason to say you did it and are responsible for it even though it was undetermined whether you would succeed. The indeterministic noise would have been an obstacle that you overcame by your effort. (Kane, 2002a, p.229)

Again, we are being asked to consider that the successful outcome of our efforts in the face of indeterminism does not undermine either our ownership of the success we achieve – we did what we were trying to do – or our responsibility for that success. Intentional action, and the responsibility attendant on success in intentional action, would appear to survive the 'intervention' of indeterministic processes that make such success undetermined.

Kane is cognisant of the fact that these three examples are not isomorphic with the case of the businesswoman, or with SFAs more generally. In particular, the case of the businesswoman requires that she have willed, and is responsible for, either outcome; whereas we would be inclined to say that if the assassin or husband failed in their respective actions, that they failed "by chance". Kane nevertheless argues that if, as in the case of the assassin and the husband, indeterminism or chance *per se* does not remove or reduce responsibility for the acts involved, then he can make a stronger case for there being will and responsibility in instances of SFAs:

> Imagine in cases of inner conflict characteristic of SFAs, like the businesswoman's, that the indeterministic noise which is providing an obstacle to her overcoming temptation is not coming from an external source, but is coming from her own will, since she also deeply desires to do the opposite. Imagine that two crossing (recurrent) neural networks are involved, each influencing the other, and representing her conflicting motivations… The two networks are connected so that the indeterministic noise which is an obstacle to her making one of the choices is coming from her desire to make the other, and vice versa – the indeterminism thus arising from a tension-creating conflict in the will… In these circumstances, when either of the pathways "wins" (i.e. reaches an

> activation threshold, which amounts to choice), it will be like your solving the mathematical problem by overcoming the background noise produced by the other. (Kane, 2002a, pp.229-30)

Kane thinks that we can say of the businesswoman's eventual action that she did it, and that she is responsible for it, whichever of the two actions is chosen. He thinks he can escape any charge that the actions of the woman are merely random, inadvertent, accidental, etc., because either act has been willed by the agent for reasons that the agent endorses; and in this sense, what is done is done on purpose.

But Kane needs to make a further claim plausible – the idea that the agent under such circumstances is in control of their actions. He concedes that, for SFAs that are undetermined, agents do not control which outcome will occur before it occurs. Yet he thinks that this does not preclude our saying that an agent controls or determines the outcome when it occurs – that is, in the moment of deciding. When agents experience a conflict of motivation characteristic of SFAs that sets up the kind of indeterministic interference described above, they can be said to have what Kane calls "plural voluntary control" over their options:

> …they are able to bring about *whichever* of the options they will, *when* they will to do so, for the reasons they will to do so, on purpose rather than accidentally or by mistake, without being coerced or compelled in doing so or willing to do so, or otherwise controlled in doing or willing to do so by any other agents or mechanisms… The conditions can be summed up by saying, as we sometimes do, that the agents can choose either way, *at will*. (Kane, 2002a, pp.230-1, italics in original)

So Kane thinks we can get intention, voluntariness, control and responsibility for undetermined SFAs because:

> …*whichever way the agents choose* they will have succeeded in doing what they were trying to do because they were simultaneously trying to make both choices, and one is going to succeed. Their failure to do one thing is not a *mere* failure, but a voluntary succeeding in doing the other… And when [they succeed] in doing one of the things [they are] trying to do, [they] will endorse that as [*their*] resolution of the conflict in [their] will, voluntarily and intentionally, not by accident or chance. (Kane, 2002a, pp.231-2, italics in original)

Kane argues that we are, in a significant sense, going up against many well entrenched habits of thought in trying to comprehend choices that are, by his own lights, undetermined. On one hand, there is an association between something involving indeterminism and it being a mere matter of luck or chance. Indeed, we can easily think of strong associations between various classic (if nevertheless debatable) examples of indeterministic processes – coin tosses, rolling of dice – and various decisions or games of chance in which such 'indeterministic' mechanisms are used. Kane wants to get leverage between such ordinary uses and

associations by insisting that ‘indeterminism’ is a technical term that is to be used simply to discriminate certain causal processes – nondeterministic or probabilistic ones – from deterministic causal processes. As applied to different causal processes, ‘indeterministic’ simply picks out processes in which the outcome of the process is not inevitable or uniquely predetermined by prior conditions. As such, it is a mistake to associate ‘indeterministic’ and ‘undetermined’ with ‘uncaused’. We do not have an absence of causation in such cases, but rather the presence of nondeterministic causation. It is causation nevertheless.

On the other hand, Kane wants to take issue with a habit of thought that seeks to impose a certain parsing of events, and an associated kind of linearity, on the hypothesized indeterministic decision he is describing. The overall image is one where the agent makes an effort to have one motivational set preside over a competing set, only for indeterminism or ‘chance’ to take over at the last minute and ‘decide’ the issue (and, we might want to add, ‘decide’ the issue at random). But Kane argues that this is a mistake, specifically in so far as the indeterminism he has in mind gets separated from the efforts of will involved:

> One must think of the effort and the indeterminism as fused; the effort *is* indeterminate and the indeterminism is a property of the effort, not something separate that occurs after or before the effort. The fact that the effort has this property of being indeterminate does not make it any less the [agent’s] *effort*… [The] whole process is [the agent’s] effort of will and it persists right up to the moment when the choice is made. There is no point at which the effort stops and chance “takes over”. (Kane, 2002a, pp.232-3)

(This is perhaps where Kane’s account has its most distinct advantage over what he labelled ‘extra factor’ theories such as agent-causal accounts – the latter libertarian accounts being inclined to ‘place’ the indeterminism at the end of a deliberative process and, esp. in the case of agent-causal theories, implicitly buy into the more compatibilist view of choice as decisive determination of the will – hence the need for an extra factor.)

Kane considers four further potential criticisms of his account. First, he asks if it might be questioned whether, because they are undetermined, we can properly consider the efforts of an agent (such as the businesswoman) as choices at all. The line of questioning could draw on a distinction between happenings and doings: can events that are undetermined be properly considered as *doings* – choices and actions – or should they rather be regarded as things that merely happen? Kane thinks this line of criticism is question-begging in that it assumes that choices and actions are things that must be determined. Instead, we should think of a choice as:

> … the formation of an intention or purpose to do something. It resolves uncertainty and indecision in the mind about what to do. Nothing in such a description implies that there could not be some indeterminism in the deliberation and neural processes of an agent preceding choice corresponding to the agent's prior uncertainty about what to do… Self-forming choices are undetermined, but not uncaused. They are caused by the agent's efforts. (Kane, 2002a, p.234)

While Kane's response must await further critical evaluation, at least the charge of begging the question seems fair. There does indeed seem to be a tendency within the compatibilist-dominated traditional debate for notions of choice and action to be unpacked in such a way that anything less than determination by the agent amounts to a move away from the domain of choice and action proper.

Which brings us to the second criticism discussed by Kane. If the above argument is sound, such that indeterminism does not undermine something's being a choice, might not the indeterminism nevertheless undermine any claim that the choice is that *of the agent*? That is, we might ask (as many compatibilists have done) whether the insertion of an element of indeterminism in the decision-making process might only serve to impinge on the extent to which we can consider the outcome of that process as a choice made and 'owned' by the agent. Kane thinks this criticism misses the mark as well. He wants to insist, in the example of the businesswoman, that her eventual choice is her own because it results from her efforts and deliberations, and these are in turn influenced causally by her reasons and intentions:

> And what makes these efforts, deliberation, reasons, and intentions *hers* is that they are embedded in a larger motivational system realised in her brain in terms of which she defines herself as a practical reasoner and actor. A choice is the agent's when it is produced intentionally by efforts, deliberation, and reasons that are part of this self-defining motivational system and when, in addition, the agent *endorses* the new intention or purpose created by the choice into that motivational system as a further purpose to guide *future* practical reasoning and action. (Kane, 2002a, pp.234-5)

Thus, ownership of choice and action stems from the relation between these and the agent's motivational system and deliberative efforts. Nowhere is it implied that this relation needs be one of determination. We have seen that, on Kane's account, the indeterminism involved in SFAs stems from the very make-up of the agents motivational set, and in that sense is an expression of who and what the agent is. And Kane wants to add a prospective view too – that agents can not only endorse a particular (undetermined) choice as their own, done intentionally for reasons, by way of expressing and justifying ownership of that choice, but they further endorse their new intention by factoring it into future deliberations. One might ask how much more evidence of ownership one could ask for than all this?

One thing that could be asked for, however, is for an explanation of how all of this relates to the important issue of *control*. The third criticism considered by Kane concerns precisely how much control an agent, such as the businesswoman, can be said to have over choice and action in a SFA. Specifically, can it not be charged that the presence of indeterminism in SFAs must, at the very least, diminish the degree of control that agents have over such actions and choices, even if it is allowed that these are choices and actions owned by the agent? For Kane (2002a, p.235), this amounts to the criticism that "indeterminism, wherever it occurs, seems to be a *hindrance* or *obstacle* to our realising our purposes and hence an obstacle to (rather than an enhancement of) our freedom."

At this point, Kane makes an interesting concession. He thinks we should concede that indeterminism does diminish control over what agents do, and that it is a hindrance to agents realising their purposes. Yet this concession is not the admission of defeat that it might otherwise appear to be. Kane, unlike most other libertarians, has argued that indeterminism need not be a general feature of deliberation and choice, but only a crucial feature of self-forming choices and actions. So, for a start, Kane's account does not imply some quite general reduction in control of agents over their choices and actions. More significantly, though, Kane sees the indeterminism arising in SFAs as arising from the agent's own will. The indeterminism does constitute an obstacle or hindrance to the achieving of a certain purpose, but it is a hindrance arising out of the motivational structure of the agent themselves in so far as they have competing and mutually incompatible purposes, each backed by their own incompatible (and typically incommensurable) sets of reasons. There is indeterminism because there is a conflict of motivation, the conflict of motivation implying that each of the conflicting purposes in some sense is standing in the way of the achievement of the other purpose. But this just seems to make the diminution of control involved in SFAs an anticipated, if not a necessary, feature of what it means for us to be capable of genuinely undetermined, self-forming actions that place our lives on fresh trajectories.

It is worth considering, in some detail, what Kane considers as the consequences of entertaining anything other than his embracing of indeterminism in the context of SFAs, again referring to his example of the businesswoman:

> If there were no such hindrance – if there were no resistance in her will – she would indeed in a sense have "complete control" over one of her options. There would be no competing motives that would stand in the way of her choosing it. But then also she would not be free to rationally and voluntarily choose the other purpose because she would have no good competing reasons to do so.

> Thus, by *being* a hindrance to the realisation of some of our purposes, indeterminism paradoxically opens up the genuine possibility of pursuing other purposes – of choosing or doing *otherwise* in accordance with, rather than against, our wills (voluntarily) and reasons (rationally). To be genuinely self-forming agents (creators of ourselves) – to have free will – there must be at times in life obstacles and hindrances in our wills of this sort that we must overcome. (Kane, 2002a, pp.235-6)

I think it difficult to underestimate the importance of this move by Kane. He has, *prima facie*, managed to side-step many of the most well-rehearsed criticisms of indeterministic accounts of free will, through a combination of making the indeterminism in SFAs an *expression* of the agent's own motivational structure, and then turning the consequences of this indeterminism (diminution of control, the presence of a hindrance or obstacle to the will) into an important feature of what it means to be free, in so far as the link between freedom and self-formation has been successfully established. And he appeals to a sensible-enough sentiment – that when we are faced with incommensurable motivations and resulting conflicts of will, we are in an important sense less in control of the direction our life will take, as compared to the level of control we might experience when we unequivocally 'know what to do'.

Kane finally turns to a fourth potential criticism, one that he considers to be perhaps the most telling against his account. The criticism is, simply, that undetermined self-forming choices are, in the end, arbitrary: "A residual arbitrariness seems to remain in all self-forming choices since the agents cannot in principle have sufficient or overriding *prior* reasons for making one option and one set of reasons prevail over the other." (Kane, 2002a, p.236) Again, Kane makes an interesting concession in acknowledging some truth to this claim; but he thinks it is a truth that reveals something important about free will and self-forming actions. Kane thinks that undetermined self-forming choices are the beginnings of what he calls 'value experiments'. As the connotations of 'value' should suggest, these choices amount to a commitment to how one's life will and should be in the future. In the absence of prior sufficient or overriding reasons at the time of making the choice, the choice is not dictated or fully prescribed by the agents past and their current psychological make up; and a more complete justification of the choice must lie in the future, depending on how the 'experiment' of setting off down the particular chosen path turns out.

Rather than side-stepping the charge of arbitrariness, Kane sees himself as returning to the conceptual and linguistic roots of the word 'arbitrary'. He reminds us that medieval philosophers referred to free will as "*liberum arbitrium voluntatis*" – free judgement of the

will. And he further reminds us of an associated notion – that of an arbiter – in a pithy summary of his view:

> …[Agents] who exercise free will are both authors of and characters in their own stories all at once. By virtue of "self-forming" judgements of the will (*arbitria voluntatis*) (SFAs), they are "arbiters" of their own lives, "making themselves" out of a past that, if they are truly free, does not limit their future pathways to one. (Kane, 2002a, p.236)

At first glance, it appears that Kane has pulled off something remarkable. He has given fresh life to indeterministic libertarianism as a position on the question of free will. He has done so without postulating any mysterious extra factors or special causal relations – indeed, by his own lights, without postulating any ontological entities or relations not already required in non-libertarian accounts of agency and free will, or elsewhere in science. And he has offered ways out, using both arguments and 'conceptual therapy', of some of the most notorious libertarian sticking points, including accusations of chance/ randomness, loss of control, and arbitrariness. The question is: does Kane's account really work?

Evaluating Kane's Libertarianism

Robert Kane's theory comprises one of the freshest contributions to the traditional philosophical debate on free will. As such, it has not escaped notice and considerable critical attention. In Kane's (2002a) article that I have used as my primary source alone, he spends the latter part of his paper considering and responding to criticisms that have been offered by what we might otherwise think of as his libertarian allies – specifically from proponents of agent-causal theories. I will, as far as possible, avoid a review of these criticisms and Kane's responses, for two reasons. First, I am sympathetic to Kane's argument that agent-causation theories and, more generally, what he calls 'extra factor' theories, tend to boil down to a problematic dualism that we should avoid if at all possible, given the well-rehearsed difficulties with dualism. Second, I think that Kane is right in trying to restrict himself, as far as possible, to an ontology and a psychology that is (or can be) shared by compatibilists and libertarians alike – at least for the purposes of generating movement in the traditional debate. Kane calls this a commitment to naturalism; and, to the extent that the term 'naturalism' can be understood in a sufficiently open-minded way, stripped of some of its more problematic associations with physicalism, I will share that commitment to naturalism for present purposes.

The more interesting sources of critical interest, then, will come from the compatibilist camp. This is especially the case because Kane has apparently been able to make important

concessions to compatibilism, notably by restricting the role of indeterministic processes to a subset of life choices – SFAs – and thereby allowing that many important life choices and/or actions, such as Luther's refusal to recant, can be understood as determined, and thus lacking in availability of real alternative possibilities. The latter example, of course, comes from Dennett, and it thus seems pertinent to turn to him to see how convincing he finds Kane's theory.

Dennett on Kane: 'A noble failure'

In his book *Freedom Evolves*, Daniel Dennett (2003) obliges by offering a critical discussion of Kane's account which, in Dennett's opinion, is the best attempt so far to construct a positive indeterministic theory of human choice. Part of the reason for Dennett's high opinion of Kane comes from a shared commitment to naturalism of the kind I have described above. Dennett frames Kane's project as follows:

> The challenge Kane faces is to describe a way our *apparent* decision-making could be *real* decision-making, and he wants to do this without postulating any supernatural entities or mysterious forms of agency. He is, like me, a naturalist, who assumes that we are creatures of the natural order whose mental activity is dependent on the operations of our brains. (Dennett, 2003, pp.102-3)

In the context of the project of Dennett's book, he thinks that he has exposed the illusion that determinism renders decision-making only *apparent* decision-making, and he thus entertains Kane's theory as something of an exercise in seeing how the mistakes of the best of the opposition can further illuminate what needs to be said about freedom from within Dennett's compatibilist framework. In discussion Dennett's critique of Kane, I will (as far as possible) try to be more open-minded about the possibility that Kane is onto something important, remain neutral on the prospects of compatibilism, and do my best to defend his account against Dennett's attack.

Dennett (2003) makes two central claims in his evaluation of Kane's theory: (i) that the indeterminism Kane wants to insert, and thus the indeterministic/ libertarian aspect of theory as a whole, is unmotivated; and (ii) that such cases of genuine indeterministic choice would, in any event, be undetectable. The criticisms under (i) go to the heart of the disagreement between a compatibilist like Dennett and an incompatibilist like Kane; those under (ii) address both the plausibility of Kane's account as well as the practical (including legal and moral) utility of the theory in making judgements of responsibility.

The crucial argument for (i) draws on what Dennett considers to be a fallacy of looking for regress-stopping tokens (objects or events) to secure our classifications of those tokens. His chosen illustration is the case of looking for a 'Prime Mammal'. Consider the following argument:

(1) Every mammal has a mammal for a mother.

(2) If there have been any mammals at all, there have been only a finite number of mammals.

(3) But if there has been even one mammal, then by (1), there have been an infinity of mammals, which contradicts (2), so there can't have been any mammals. It's a contradiction in terms.

(Dennett, 2003, p.126)

The problem in the argument seems clear enough. We don't need to be able to draw a distinct line in the ancestry of mammals in order for it to be the case that there are, and have been, undeniable cases of mammals. On Dennett's diagnosis, the problem lies with the pursuit of a regression-stopping case; and he points out that the illness often leads to something equally troublesome – essentialism. You don't need a (and, according to Dennett, evolutionary biology shows us that there is no) list of essential features for something to be a mammal in order that there are mammals. So far so good. I think Dennett is right about this as a problematic tendency within (and outside of) philosophy, and I share his anti-essentialist views on the matter.

How does this argument relate to Kane's account of free will? Dennett correctly notes that Kane has set off in pursuit of regress-stopping choices and actions in the lives of human agents. He directs us to the following quote from Kane:

> If an infinite regress is to be avoided, there must be actions somewhere in the agent's life history for which the agent's predominant motives and the will on which the agent acts were *not already set one way*. (Kane, 1996, p.114)

As we have already seen, this is the means by which Kane thinks we can locate ultimate responsibility within an agent: postulating the existence of regress-stopping SFAs in the life history of agents. If we did not find these undetermined SFAs – if we were to find sufficient cause for all an agents actions and choices – we would be led endlessly backwards towards events that predate the existence of the agent, and this threatening regress would undermine any claim to be, in Kane's terms, the "ultimate creator and sustainer of one's own ends and purposes" (Kane, 2002a, p223).

However, we cannot find fault with Kane just because he has argued for the importance of finding buck- and regress-stopping events in order to secure ultimate responsibility for agents. Looking for a prime mammal might be wrong, even absurd, but that surely does not

render problematic all attempts at regress-stopping? If I claim citizenship of a country based on my ancestry, my claim will rest on various facts about my ancestors such as when and where they were born, whether, when and where they were married, when and how they were naturalised as citizens of the country of my birth. I could not contest a refusal to grant me citizenship on the grounds that the authorities are complicit in committing a logical fallacy akin to searching for a prime mammal. And enough people have been concerned about the regress that determinism appears to give rise to for us to think that this might be a case of a regress worth stopping.

Much of Dennett's (2003) discussion is dedicated to demonstrating how, by his lights, Kane's account won't work. That is, he pays most of his attention to the epistemic and practical difficulties of detecting SFAs – criticisms I have chosen to segregate under (ii) (above) and will discuss in due course. When he eventually addresses Kane's central concern (which derives from an incompatibilist conviction about events needing to be up to the agent in a way that determinism does not seem to allow), Dennett sums up his criticism as follows (referring to a simple example of a choice to go or stay):

> We can now recognise that [the incompatibilist argument] commits the same error as the fallacious argument about the impossibility of mammals. Events in the *distant* past were indeed not "up to me," but my choice now to Go or Stay is up to me because its "parents" – some events in the *recent* past, such as the choices I have recently made – were up to me (because *their* "parents" were up to me), and so on, not to infinity, but far enough back to give my *self* enough spread in space and time so that there is a *me* for my decisions to be up to! The reality of a moral me is no more put in doubt by the incompatibilist argument than is the reality of mammals. (Dennett, 2003, pp.135-6)

Whilst this may sound sensible enough as a statement of a compatibilist position on what it means for my decisions to be up to me, it is not immediately clear exactly how Kane's argument has been rebutted. Compatibilists like Dennett may be willing to live with varieties of freedom (and responsibility) that are compatible with determinism, and they are consequently charged with producing an account of the circumstances under which we can sensibly talk about choices being 'up to me' in some relevant sense. Kane, however, might be unmoved by all of this. Indeed, it sounds a lot like the stalemate I described in over-viewing the state of play in the traditional philosophical debate about free will. Incompatibilists want one kind of responsibility; compatibilists deny that it is available, and make do with something else. Merely pointing to the good reasons that exist for avoiding excessive regress-halting (and essentialism) doesn't establish that Kane is wrong to want to stop this particular regress.

It seems to me, then, that Dennett's criticisms under (i) – that Kane's resort to indeterminism is unmotivated – cannot really be separated from (ii) in which he wants to argue that the kinds of indeterministic choices Kane offers us are, ultimately, undetectable. We can, I think, see this in Dennett's explanation as to just what we should be suspicious of in Kane's positing of SFAs:

> …one should be suspicious of the demand that there be an event – an SFA – that has some special, intrinsic, local feature that sets it apart from its nearest kin and explains its capacity to found something important. Is it plausible that an agent who hadn't yet experienced one or more of these very special events (but only near misses, pseudo-SFAs) would simply not be responsible for any acts performed? (Dennett, 2003, p.128)

This quote follows a comparison, suggested to Dennett by Paul Oppenheim, between SFAs and so-called speciation events in evolution. Since every birth involves small differences in offspring that make them different, and every difference could turn out to be something that eventually shapes a new species, every birth is potentially a speciation event. But speciation events can only be identified retrospectively, and we should not look for anything special at the time of a birth that will turn out to have been a speciation event.

It would seem that the notion of parentage is doing a lot of work in Dennett's assessment of Kane's SFAs. As we have seen, Dennett's brand of compatibilism favours an accumulation of ownership and responsibility for choices over time, a building up of choices that are 'up to me'. In contrast, Dennett wants to suggest there is something suspicious, if not something risky, about specifying and trying to identify special events in the life of an agent whose presence will be the grounds of the responsibility we can attribute to them. We are invited to share his view, expressed in his rhetorical question, that it is implausible to think that *only* agents who had experienced one or more undetermined SFAs would be held responsible for the actions they perform. As an empirical claim, of course, this seems right – we certainly don't go searching for SFAs (as defined by Kane) when attributing responsibility. Nor do we descend into the murky depths of quantum physics to settle such questions by testing for the requisite indeterminism in the brain.

Yet Kane, as we know, is aware of the fact that his theory requires some degree of empirical proof, at least to the point of finding evidence of indeterministic and chaotic activity of the right kind in the functioning of the brain. He is in the same boat as any other indeterminist – proponents of agent-causal theories, Searle, etc. – who require that it turn out to be the case,

empirically, that the brain's functioning is not entirely deterministic. Where Kane has some advantage over these other accounts is that he requires *less* indeterminism; and he can plausibly allow for determined 'considered' life choices (Luther) as well as for snap decisions determined by our settled dispositions while retaining indeterminism as the grounds for the responsibility we attach to such choices. Dennett thinks there is a fundamental problem with the role Kane wants SFAs to play, but it is not yet clear what that problem is, beyond the obvious point that compatibilists will (obviously) go about the exercise of assigning responsibility in a different way.

We can get closer to the issue if we look at Dennett's attempt to examine the case of Luther through the lens of Kane's theory. Kane thinks that Luther can only be ultimately responsible for his (determined) refusal to recant if there were choices and actions in his history where he could really have done otherwise. Dennett asks us to then consider our carrying out the requisite study of Luther's life history, and makes the following claims:

> …nothing we could discover about such macroscopic details would shed *any light at all* on the question of whether or not Luther had had any genuine SFAs during this period. We could certainly discover that episodes of conflict and soul-searching occurred on various occasions, and we might even confirm that these occasions set up "chaotic" opponent processes in the neural networks from which his decisions eventually emerged. What we could not discover, however, was whether these tugs-of-war had the benefit of genuinely random, as opposed to mere pseudo-random, sources of variability. The price libertarians must pay for sequestering their pivotal moments in subatomic transactions in some privileged place in the brain (at time *t*) is that they render these all-important pivots undetectable by both the everyday biographer and the fully equipped cognitive neuroscientist. (Dennett, 2003, pp.128-9)

Dennett's central argument in this passage appears to be an epistemological one: he thinks there is nothing that biographers or the most well-equipped neuroscientists could do to find the indeterministic events Kane posits as underpinning the undetermined outcomes of SFAs. Yet on the one hand, Kane's claim (as Dennett well knows) is both metaphysical and empirical. It is not clear how Dennett's argument addresses the metaphysical claim, and it is not exactly self-evident how Dennett can be so forthright in his dismissal of the empirical claim. Why is it so obviously problematic for indeterminists like Kane (or Searle) to think that neuroscience can be charged with the empirical task of discovering genuine indeterminism in the brain processes underpinning undetermined free choices?

The answer seems to lie in Dennett's views on randomness, and his discussion of random versus pseudo-random processes. To put the point somewhat bluntly, Dennett has an axe to grind. When he hears talk about 'special powers' and 'special cases', he puts his engineer and

cognitive scientist hats on and proceeds with a lecture or two about how, if people only properly understood notions like algorithms, chaos, quantum randomness, pseudo-randomness, and neural networks (to name a few), we wouldn't land up in all the kinds of trouble that Kane (and others) would have us get into. Indeed, his discussion of Kane's theory is peppered with mini-lectures on each of these concepts, yet in such a way as to not always make it clear exactly what it is about *Kane's* theory Dennett finds so objectionable. I will do my best to extricate the argument as it applies to Kane from all the rest.

From a functional, engineering point of view, there may be any number of reasons for wanting to insert a degree of randomness into a given process. Dennett (2003) refers to the example of computer programming where, at various points, an application might 'call out for' a random input; and this is typically done through a 'request' to a random number generator. But – and Dennett wants to make maximum mileage from this – in the case of computers and their random number generators, what we have is at best *pseudo*-randomness. The random number generator will itself be an algorithmic process reliably and deterministically designed to churn out what are, for most purposes, and specifically in relation to the application calling for a random input, random numbers.

The deterministic algorithm underpinning such a procedure makes this merely pseudo-random. Yet, from the point of view of our engineer and her application, that is good enough. Moreover, it would make no meaningful difference if the system were to be enhanced by hooking up the random number generator to a genuinely random system, such as a Geiger counter in proximity to a radioactive substance. From a functional point of view, this would be (to use one of Dennett's favourite phrases) a difference that makes no difference.

Dennett (2003) thinks that the above observations about systems with random and pseudo-random number generators can be used to generate two problems for Kane. The first relates to the apparent epistemological argument we saw earlier. It seems Dennett thinks that the argument is not merely epistemological – that is, it is not merely an empirical speculation concerning the likelihood of biographers and neuroscientists ever being able to find the indeterministic events that Kane needs. Rather, Dennett appears to be making a more principled claim to the effect that, because (from a functional point of view) the difference between a random process and a pseudo-random process in Kane's faculty of practical reason is a difference that makes no difference, we *could not be* in a position to distinguish

genuinely indeterministic SFAs from life choices whose apparent indeterminism is only underpinned by pseudo-random processes (and that therefore fall within, or are at least compatible with being a part of, a larger deterministic system). Thus, the difference that makes no difference (at the level of neural events underpinning the machinations of the faculty of practical reason) translates, by Dennett's logic, into a difference that makes no difference when it comes to distinguishing agents and their choices. If it is impossible to distinguish genuine, undetermined SFAs from merely pseudo-indeterministic (because pseudo-random) life choices, then we are doomed on Kane's theory to not being able to distinguish free and responsible agents from those who only closely resemble such agents, but who do not (by Kane's lights) deserve praise or blame because *all* their behaviour is in fact determined.

The second problem Dennett wants to put forward for Kane is one concerning boundary problems. Dennett discusses a number of these 'boundary problems' in his chapter on Kane, and it is not always clear exactly to what extent Kane need worry about all the alleged boundary problems raised. But, in terms of the apparent overall structure of Dennett's argument, there is at least one boundary problem that might be telling against Kane's account. From an engineering (or cognitive engineering) perspective, Dennett holds that there are any number of ways that the randomness Kane wants in his faculty of practical reason might be 'designed into' the functional space of that faculty. There could be randomness in the *inputs* to the faculty, there could be a random process *within* the workings of the faculty, or the faculty could 'send out' for a random factor at certain points in certain kinds of processes within the faculty. Glossing over Dennett's more detailed discussions of each of these possibilities, it seems that the overall point against Kane is that much of this would be arbitrary line drawing in functional space (as Dennett (2003, p.122 ff) likes to put it, "If you make yourself really small, you can externalise virtually everything"). Most specifically, Dennett wants to claim that, because *randomness is just randomness*, no matter where you get it from, it makes no great difference if the randomness Kane wants within his SFAs comes from chaotic amplifications of quantum indeterminacies *arising from* neural processes in the brain, or from remote recordings of quantum events in a radioactive substance as recorded by a Geiger counter, and transmitted to the brain.

In summary, Dennett thinks that Kane's account flounders on the same issues concerning randomness that are the mainstay of compatibilist critiques of indeterministic libertarianism

in general. On one hand, a functional system with a chance factor built in is insensitive to our distinctions between real randomness and pseudo-randomness: "…chance looks exactly the same, whether it is genuinely indeterministic or merely pseudo-random or chaotic." (Dennett, 2003, p132). And, on the other hand, where the randomness comes from, and specifically whether it is internal to the agent and their neural processes or not, makes no difference to the role played by that randomness: "Randomness is just randomness; it isn't *creeping* randomness." (Dennett, 2003, p.133).

Evaluating Dennett

Dennett (2003) thinks of Kane's theory as something of a noble failure – the best attempt to work out a defensible incompatibilist theory of free will that is helpfully revealing in its failure. Yet, as Dennett's own references to his influential *Elbow Room* imply, much of his critique is based on points and arguments he raised in 1984. Avid historian of the free will debate that he is, Kane is clearly familiar with Dennett's arguments in *Elbow Room*. Has the libertarian Kane really failed so dismally in moving the libertarian side of the debate beyond the barricades put up by the compatibilist Dennett?

Dennett's (2003) most significant concession to Kane is that he has made a significant contribution to the debate by focussing our attention on the issue of 'plural rationality' – Kane's idea that, at the heart of SFAs, there lie conflicts between sets of reasons that we own as ours, such that whatever the outcome, it is one that we endorse as ours done on purpose for our own reasons. Dennett thinks that this kind of scenario has been largely ignored within the traditional free will debate, and that Kane makes a convincing case for moving these kinds of choice situations to centre stage. But he also thinks that the account of free will built on this notion of plural rationality fails because (a) it doesn't need the ingredient of indeterminism that Kane wants to add to the recipe, and (b) because it can't harness the posited indeterminism in a way that would make the resulting character-building distinguishable from determinism. These claims are clearly related, and I have done my best to reconstruct the central arguments offered by Dennett to support them.

What Dennett does not offer us, however, is a case-specific argument to the effect that Kane (or any other libertarian) is wrong to want to stop the particular regress that, if determinism is true, appears to make the 'causal parents' of an agents actions ultimately lie in events that predate their existence. The comparison to the fallacy of looking for prime mammals is

useful, and probably persuasive if one is already of compatibilist inclinations, but I have already argued that it cannot be used across the board as an argument against regress-stopping. In addition, there is a similarity of approach between Kane's work and the accounts of free will put forward by many compatibilists that is obscured by an excessively dismissive attitude towards regress-stopping.

Dennett wants responsibility to 'accumulate' in agents, not by finding regress-stopping special life choices such as Kane's SFAs, but through a sufficiently spread out succession of choices and actions whose causal parents were themselves choices and actions that were meaningfully 'up to' the agent. Kane also wants responsibility to accumulate through a succession of important and difficult life choices – his character-building SFAs. As I have noted, this feature of Kane's theory reduces the frequency with which indeterminism is called upon to secure a libertarian form of ultimate responsibility whilst leaving the account better able to deal with life choices that are determined, whether these be Luther-like stands of principle or the snap decisions of everyday life (that are not plausibly preceded by deliberation or indeterministic vertigo of choice). But if Kane's arguments succeed, then he, unlike Dennett, can offer his SFAs as regress- and buck-stopping moments of choice that offer a degree of ownership (and responsibility) to the agent that compatibilism cannot. There is no possibility of absconding from responsibility because conditions in the universe, plus the laws of nature, were sufficient to make it the case that the agent committed some reprehensible deed, because there will be character-shaping SFAs in their life history for which there were no sufficient antecedent causal conditions. Moreover, in the absence of such sufficient antecedents, it looks like the best explanation of the outcomes in such SFAs lies in the conflicting motivations of the agent, and in the reasons they had for choosing and acting in the manner they eventually did, because Kane has tried to show us how such explanations (and associated attributions of responsibility) can survive the intervention of indeterminism (in the form of chaotically amplified quantum indeterminacies) in the shaping of these outcomes. Surely we might think that this confers a form and degree of ownership of our choices and actions worth wanting?

The crux of the matter seems to be one of ownership. Compatibilists (like Dennett) need to tell a plausible story about ownership in a context where any given choice or outcome will, if determinism is true, have had sufficient causal antecendents. This can be seen as raising problems as to whether there was any openness of possibilities in terms of outcomes, and the

potential for a steady regress back to a time and conditions predating the existence of the agent. Libertarians have traditionally tried to accommodate alternative possibilities and regress-stopping ownership by inserting a variety of indeterminism within the agent, only to have compatibilists (and hard determinists) point out that randomness and indeterminism seem to *weaken* claims of control over alternative outcomes, and ownership of outcomes. Kane tries to sidestep this by making the agent both own the indeterminism – it results from the operation of competing motivational sets within the agent – and the results of the choice involved – by showing how we can act for reasons, on purpose, despite the indeterminism involved in SFAs.

Perhaps the best way for Kane to respond to Dennett is by emphasising the origin and ownership of the indeterminism he posits in SFAs, and tackling head-on the accusation that "randomness is just randomness." It matters to Kane that the indeterminism he posits is, in some sense, inside the faculty of practical reason, between input and output (as Dennett puts it), because this is not an exercise in cognitive engineering or artificial intelligence, but a philosophically motivated empirical hypothesis about the nature, structure and function of human deliberation under conditions of conflicting motivation. Whether or not we can design some 'cognitive' artefact (like a computer program) that could mimic the set-up Kane describes by 'sending out' for a random or pseudo-random number at the point where quantum indeterminism is being hypothesised as playing a role in the human brain's realisation of SFAs cannot be a decisive point against Kane (unless, like Dennett, you are already inclined to think that just about everything is algorithmic!). In other words, much of Dennett's discussion concerning various alleged boundary problems raised by Kane's chosen means for securing ultimate responsibility makes it sound like it is arbitrary where, and from what source, you insert the required randomness. But Kane thinks this is far from arbitrary – the indeterminacy is there *because* the agent has competing reasons, and *because* those reasons are not decisive in resolving the conflict one way or another. This is a sense in which the agent *owns* the indeterminacy that is in their will in a way that they could not own the indeterminacy of a radioactive-substance-plus-Geiger-counter connected by remote to their brain. And the indeterminism arises when and where it does, during deliberation, between input to and output from the faculty of practical reason, *because* it is under these conditions of conflict between competing and incommensurable reasons that Kane hypothesises the chaotic amplification of quantum indeterminacies to a level where they can affect the output of that faculty.

What of the charge that “randomness is just randomness”? We can recall that Kane has already conceded an important sense in which the account he is offering does imply a certain loss or reduction of control, and it was my expressed view that this concession to compatibilist critics of libertarianism was a significant one – it promises to reduce the sting of the criticism that inserting indeterminism and randomness in the will makes our will less our own, and our actions less like something we do and more like something that happens to us. Instead, for Kane, the randomness is there because of our motivational set-up. Its being there and, in the case of SFAs, its role in our choosing one way or the other is a function of the conflict that is internal to our will that gets resolved within our will when we choose one of the available ways forward. We might, from Kane’s point of view, wish to contrast this with our *choosing to* resolve a conflict in our will by making our decision depend on the outcome of some (external) random or pseudo-random process – a coin toss, a reading from a Geiger counter in a given time interval, the selection of a certain number in a lottery or by a computer’s random (or, for Dennett’s sake, pseudo-random) number generator. Here, our choice is to make our decision be *determined* by something else (and, of course, to further choose to make that something else in fact determine our will when the appropriate time comes, unlike the gambler who undertakes to stop if a significant win has not happened in a given space of time, only to revise the allotted time interval when the win has not transpired); in Kane’s SFAs, our choice is the undetermined outcome of our deliberations. Randomness plays a role either way, but surely Kane can insist that the role is different in the case of SFAs. The boundaries of internal versus external are not arbitrary, and the randomness is not just any randomness.

Furthermore, Kane might well respond that Dennett’s expressed view that “randomness is just randomness” implicitly depends for any claimed obviousness on a fallacious equating of some thing’s being undetermined with it being uncaused. Take any case of what is generally considered to be probabilistic causation rather than (what is often at best only assumed to be) deterministic causation. To the extent that anything is determined in a probabilistic causal set-up of some kind, it is the probabilities of the various possible outcomes. What is not determined is the eventual outcome. In the case of quantum mechanics, as I understand the generally held interpretation of the theory in this branch of physics, there are no gap-filling other or hidden variables that intercede between the possible outcomes and their respective probabilities, on the one hand, and the actual outcome on the other hand. Which possibility

becomes actual is, in this sense, random – a matter of chance. And the 'intervention' of chance or randomness does not make the actual outcome *uncaused*. Rather, the causal influence here was non-deterministic and probabilistic – the causal factors at play made certain outcomes (but not others) possible, each with a certain degree of probability.

If we apply this kind of model to Kane's SFAs, we get a picture in which it clearly matters that the possibilities, probabilities and indeterminacy are internal to the system he is describing. It is a feature of the system that the motivational causal influences at play in SFAs are only probabilistic, not determinative. There is still causal influence here – indeed, there is causation. And there is no 'call out' for a random input or influence, any more than a quantum phenomenon 'calls out' for the intervention of randomness – the role played by chance is internal to the dynamics of the system. Of course, we must grant that the chance factor in quantum phenomena is something that is (theoretically) internal to the quantum domain, whereas Kane is postulating that amplified quantum fluctuations play a role in neural and psychological phenomena at very different levels of ontological and explanatory complexity. Yet Kane sees these quantum fluctuations as internal in so far as they are a part of his hypothesised neurophysiology of decision-making in SFAs. Randomness, it seems, is not just randomness. Chance can be, in an important sense, internal to a system, and its having a role to play in the first place can equally be an internal feature of that system. I am inclined to think that Kane has learnt the lessons Dennett (1984) offered in *Elbow Room*, and that Dennett (2003) has failed (or refused) to grapple with the subtleties and complexities of Kane's theory in this regard. We would do well, as Dennett encourages, to be critically suspicious of claims regarding special powers, special events, clear line-drawing where no clear lines need exist, or attempts to harness the quirks and mysteries of quantum physics in solving puzzles at entirely different levels of material organisation and complexity. We should not, however, take this attitude to an extreme in which all such endeavours are dismissed because the theoretical moves they try to make have been ruled out of court in advance.

More Pressing Problems

My discussion and evaluation of Dennett's critique of Kane suggests that Kane has done well to avoid the usual traps that compatibilists have set for libertarians. Moreover, I have along the way left it as at least an open question as to whether Kane is justified in going off in pursuit of Ultimate Responsibility and his regress- and buck-stopping, undetermined SFAs.

Should we, then, accept Kane's theory as a philosophically viable, if empirically unconfirmed, account of free will? Should we redraw the lines of allegiance and orthodoxy within the free will debate, and pursue an invigorated libertarian agenda?

My answers to these questions must be negative, at least as far as Kane's theory goes. For all the interest in, and sophistication of, Kane's account, I think it fails to make a convincing case for his brand of indeterministic libertarianism. Like Dennett, I think the failure is an instructive one, but for reasons that Dennett does not succeed in highlighting in his critique of Kane.

Dennett (2003) is on his firmest ground when he argues that Kane will have difficulties in distinguishing between responsible agents with real SFAs in their history from agents who, because lacking real SFAs, should not properly be held responsible. But the problem I see here is not the one emphasised by Dennett. He claims that nothing a biographer or neuroscientist could do could ever reveal the real indeterminism Kane requires for a choice to qualify as a real SFA. I would allow that, as a matter of empirical hypothesis, it is in principle possible for advances in the neurosciences (and, no doubt, other scientific domains) to make it such that we could, first, determine whether or not the brain is influenced by truly indeterministic factors and, second, be in a position to detect such indeterminism if it is there to be found. Dennett's claims as to the functional indistinctness of truly random and merely pseudo-random processes might point us to certain epistemological obstacles that will need to be overcome, but I have not seen his argument to show that this is not possible in principle.

Where I agree with Dennett is that Kane's account seems to misdirect our focus when it comes to ordinary (not to mention legally or politically pressing) judgements of responsibility. On Kane's own account, the indeterminism that he offers to secure UR is not a feature of everyday choices and actions, but only a characteristic of a potentially small set of special SFAs. That seems to mean that, in making judgements of freedom and responsibility, we are first going to have to tell something very much like a compatibilist story about the ways in which a particular (potentially determined) choice and action came about before questions of ultimate responsibility even get on the table. Kane might not be alone amongst libertarians in thinking that some of the things emphasised by compatibilists – the absence of coercion or brainwashing, for example – are obviously important to judgements of freedom and responsibility. But other brands of libertarianism are likely to be less parasitic on

compatibilist accounts because they see some form of freedom-granting indeterminism as a feature of most if not all free-choice situations. Kane, it seems, must depend much more heavily on compatibilist criteria of freedom and responsibility – criteria he then supplements with his additional condition on responsibility, namely the ultimate responsibility that comes from having genuine SFAs in one's causal history of choice and action.

While this criticism, on its own, might not sound too severe, it is reasonably straightforward to raise the stakes in the argument. For, if SFAs are allowed to be relatively rare and causally distant events in an agent's history of choice and action, then we might wonder just how much work these special cases are really doing in rendering a particular choice one that is free and worthy of judgements of responsibility. Specifically, it seems that Kane will have to face many of the puzzles and problems faced by compatibilists, in terms of how 'deep' and/or how far back responsibility must go for the various causal influences in a particular instance of choice, and then further locate SFAs within this picture so as to add the necessary ingredient of ultimate responsibility.

Let me try to illustrate what I have in mind. On a traditional account of reasons as causes, reasons are comprised of beliefs and desires. For compatibilists, an agent's choice and action on the basis of a particular reason or set of reasons will be free if we can tell the right kind of story about how the agent came to have that/those reason/s[52]. Luther will be free in his (determined) choice not to recant if we can tell the right kind of story about how he came to have the beliefs and desires that led him to choose in this way. Kane needs to tell a story like this too, and then add the ingredient of some temporally distributed SFAs that will, presumably, have relevant links to at least some of the beliefs and desires causally implicated in Luther's choice. But what those links will be, and the extent to which they touch on the particular reasons for which Luther is acting, is not at all clear. If Luther has genuine SFAs in his biography, then *any* choice he makes in the wake of such SFAs will not have causally sufficient antecedent conditions at any time in his history prior to those SFAs. Yet we are not interested in some generic form of responsibility (ultimate or otherwise) that can be conferred by making Luther's life story one that would not unfold in a deterministically inevitable way as a result of various indeterministic processes. We want to know if, how and why Luther is

[52] At least, this is true for historical/genetic compatibilist accounts. Whatever a structural compatibilist account might have to say about the causal role of reasons, the status of the choice and action as free is not tied in this way to a story about the origins of the agent's reasons. See Chapter 3 for more on the difference between historical/ genetic and structural compatibilist accounts.

making the particular choice he is making. The challenge for Luther's biographer is not, in this sense, Dennett's challenge of how one could detect real SFAs in Luther's development, but rather the problem of relating the alleged (generic) ultimate responsibility we might get from these special events to the (specific) causal antecedents of his determined choice not to recant.

In short, it strikes me that Kane has to confront two pressing issues. On the one hand, he does not appear to be exempt from the compatibilist's need to tell a story about responsibility - that is, a story that can confer enough responsibility for enough of the causal antecedents to a particular choice stretching far enough back in time. On the other hand, he must add to this a story of how the ultimate responsibility he thinks is secured by having real SFAs in one's history can be appropriately 'hooked up' with the above story, such that this ultimate responsibility can link to enough of the elements in the other story to make it seem relevant to the particular choice and action under consideration. If the latter story cannot be told, or it is told in a way that renders the links too tenuous and peripheral to the causal (reason) complex being examined, then Kane's SFAs (if they exist) would make humans metaphysically interesting without telling us much about freedom and responsibility for particular choices and actions.

There is, however, a deeper problem in Kane's account, one that might be shared by any similar indeterminist libertarian account that is not of the 'extra factor' variety. Kane thinks we get UR because SFAs are undetermined: motivational conflicts in an agent set up chaotic dynamics in the brain that amplify quantum indeterminacies, and leave the outcome of such conflicts undetermined by prior causal conditions. But an argument similar in sentiment to the one just employed in the case of particular choices and actions can be put forward as a means to undermining the significance of SFAs. I have defended Kane (against Dennett) in emphasising the ownership of the indeterminism invoked in his account that comes from the fact that the motivational conflict giving rise to a SFA is something internal to, and owned by, the agent. At least, this sense of ownership seemed sufficient to counter Dennett's claim that randomness is just randomness. But we could well ask about the ownership of the motivational conflict, and hence the SFA, in a different sense. How, we might ask, did the agent come to be in the predicament of a particular motivational conflict between incommensurable goals and reasons?

It seems that the answer to this question is, again, going to require something much more like a compatibilist account of responsibility for reasons (beliefs, desires, etc.) in order for us to feel that the conflict is sufficiently 'owned' by the agent, such that the outcome of the conflict (undetermined or not) is a meaningful expression of the agent's will. But that makes Kane's account once again parasitic on a more compatibilist understanding of issues of responsibility for reasons and motivations while rendering the hypothesised indeterminism of SFAs something of a side-issue – a probabilistic, indeterministic 'hiccup' in an otherwise (potentially) deterministic story about how people come to have the motivational set-ups they in fact have.

To illustrate the problem I am alluding to, we might consider an agent faced with a motivational conflict arising out of incommensurable goals and conflicting reasons for pursuing those goals – the basic set-up of one of Kane's SFAs. Suppose our agent suffers from obsessive-compulsive disorder, and is under the sway of two conflicting compulsions: they need to wash their hands (they have a hand-washing compulsion), but they are also compelled to check their water-mains inlet to look for any signs of leaks or sabotage, lest the water be contaminated in any way. Our agent usually arranges their living space such that the former action can readily be preceded by the latter, but today they find their path to the water mains obstructed by dirty garden tools. No-one is around to assist them, and they cannot move the tools without immediately washing their hands thereafter. Of course, they can't do this without having first checked the water-mains. Something is going to have to give.

Let us further suppose that the resulting conflict gives rise to the kind of chaotic amplification of quantum indeterminacies in the brain that Kane posits at the heart of SFAs, and our unfortunate agent makes their choice – they simply wash their hands. Is this an SFA? On Kane's account, it would appear that it is not. Kane wants to build in the same kinds of safe-guards against compulsions, coercion, brain-washing, etc. that are the familiar stuff of compatibilist accounts. The indeterminism he posits, and the fact that choices flowing from such indeterministic processes are undetermined, is not sufficient to make a choice or action into an SFA. But the example encourages us to notice two things. First, as I have been arguing above, Kane's account is clearly parasitic on a more-or-less compatibilist account to make these kinds of distinctions. Second, and perhaps more noteworthy, the choice situation just described has all the regress-stopping features of any genuine SFA with respect to the causal story that can be told about it – the antecedent causal conditions are not sufficient to

explain the outcome of the choice. So it seems that regress-stopping and buck-stopping come apart in this instance.

This might not seem too problematic. Kane is surely entitled to be as parasitic as he likes when it comes to compatibilist accounts of responsibility, and his account would surely be at fault if it allowed the undetermined actions of an OCD sufferer to qualify as freely chosen. Nevertheless, causal regress-stopping was supposed to be the key feature of Kane's account of free will, and if we continue to manipulate our examples, it becomes ever clearer that it will do very little, if any, significant work for him.

Imagine an agent with alleged genetically-inherited depressive tendencies.[53] We might easily construct a scenario in which the agent is faced with a motivational conflict with all the hallmarks of an SFA, and that issues in an undetermined choice as a result of indeterministic processes in the brain. Suppose the agent is conflicted over whether or not they should apply for a prestigious scholarship. They have strong and otherwise convincing reasons for wanting the scholarship, and believing that, at least on paper, they meet or exceed the minimum criteria for eligibility. On the other hand, their depressive personality inclines them to believe that they are unlikely to get the scholarship, and that they will thus needlessly endure the effort, stress, and ultimate humiliation and disappointment of failure if they make the application. Again, something will have to give.

Imagine that they choose either option – making the application or not. Is this an SFA? It has the requisite regress-stopping indeterminism arising from a conflict of incommensurable reasons/ motivations. It has one set of reasons – those in favour of applying – that is lacking in any obvious features (coercion, compulsion, brain-washing, duress) that would render responsibility problematic. It has another set of reasons that, though not irrational in any obvious way, are ultimately grounded in genetically-inherited depressive tendencies in our agent. Absent these tendencies and our agent would either not have these reasons for not applying, or they would be grounded in some other way (such as, perhaps, a string of relevant and similar actual incidents of applications that led to disappointment).

[53] Whether depression is genetically inherited, and the sense in which it could be so inherited, is not especially germane to the example, other than in the ways discussed in what follows.

My contention is that, in deciding on this case, the real indeterminism we have allowed as a feature of the choice situation does no significant work at all. Kane, at least, thinks it is a necessary feature if the choice is to be an ultimate-responsibility-securing SFA, but that is all it can be. It is of no use at all in deciding whether our agent has the right kind of responsibility for each of the conflicting reason sets they have, and that is surely where the important work of understanding and judging matters of freedom and responsibility is likely to lie in an account centred on reasons and rationality. Stopping a regress of sufficient causes does not give you ownership, or responsibility, or freedom.

Perhaps Kane can live with this. As already noted, he never claimed that the indeterminism of SFAs was sufficient to secure ultimate responsibility. Yet, if it secures so little in increasingly realistic and difficult cases, we must surely wonder what it can hope to *add*. Suppose our depressive agent chooses to apply for the scholarship, and we (or Kane) can offer a story about the agent on which this choice does qualify as a SFA. Perhaps we satisfy ourselves that the agent is sufficiently aware of, and has critically examined and accepted, their depressive outlook on life – they value the reduced risk of disappointment that it offers, and are willing to trade this off against some opportunities they might miss as a consequence[54]. Now we compare our depressive agent, with their genuine SFA, to a very similar agent faced by the same choice, but who finds in a recent success the decisive reason that leads them to also choose to apply for the scholarship.[55] As described, the choice cannot qualify as a SFA. It lacks the requisite unresolvable conflict and consequent indeterministic choice. One set of reasons proved decisive, and the choice was consequently as determined as any non-SFA choices are likely to be. Kane wants us to believe that there is something special about the choice of our first depressive agent that is missing in the case of the second agent.

But the differences between the two agents are, for the most part, circumstantial. As a matter of contingent fact, the second agent had experienced a recent success of some kind that featured prominently in their deliberations about the current choice. This difference of circumstance allowed them to resolve their conflict of motivation in a decisive (‘determined’) manner, while the first agent was left with an unresolved conflict issuing in an undetermined

[54] That is, we allow or emphasise a more structural account of the ownership and endorsement of these depressive characteristics in this case, such that the genetic origins of the dispositions are of less significance to the case.

[55] At the risk of stretching credibility, we could imagine these agents as identical twins, each contemplating an application for (separate) prestigious scholarships – separate simply so that the prospect of competition with their sibling does not enter the example.

choice. It is stretching credibility to imagine that the choice of the first agent is self-forming in some important way that is somehow missing from the choice of the second agent. This is especially so in the light of my preceding argument – namely that the important work in deciding that the first agent had a genuine SFA lay in telling a story about their depressive tendencies, and that story is common to the choices of both our agents. The implication was that ownership and responsibility were not secured because of the presence of indeterminism in the resolution of the conflict. By the same logic, all that we have said about ownership and responsibility for the first agent must apply to the second agent. So, why the difference in significance of the ultimate choices?

Kane could, of course, respond by claiming that the story we told about the first agent's depressive tendencies would need to posit various other SFAs along the way, and so our second agent would be free and responsible enough for their current (determined) choice in the light of this history of similar SFAs. Responsibility is accumulated over time and, consequently, we should not expect any particular SFA to look all that special, even in comparison to very similar determined choices. But now I think we can confront Kane with a dilemma. If individual SFAs have regress- *and* buck-stopping significance, then we need an explanation of why the indeterminism involved in various candidate SFAs seems so *insignificant*. We need an explanation of why individual SFAs are special when they do not appear to be all that special in isolation, and without this explanation shifting so much emphasis onto the indeterminism of the SFA as to make it the case that our OSD sufferer might count as having a SFA. On the other hand, if responsibility can only be accumulated, we are headed towards a different kind of regress – a regress of SFAs that will, at some point, most likely land up with a small set of candidate SFAs that look very much like our OCD case because we cannot tell a plausible story about ownership and responsibility for the reasons that gave rise to the motivational conflict. But then explaining the significance of these cases will require exactly the kind of explanation demanded for the significance of *any* particular SFA.

Of the options available, it strikes me that the second is the much more plausible and promising one – responsibility should accumulate over time, without too much significance being placed on any single life choice or action. But then we really are dealing with something much more like traditional compatibilism, and Kane's indeterministic hypothesis is at best an empirical speculation about ways in which quantum phenomena and chaotic

dynamics in the brain can throw up ripples of indeterminism to disrupt what might otherwise be the remarkable determining causal powers of the human mind-brain-body system. Kane's arguments may provide us with reasons to think that we can still talk of causation and responsibility in cases where such hypothesized indeterminism rears its head, and this could be a contribution to a significant bit of conceptual therapy for us if the universe (or, more locally, the brain-body system) turns out to be indeterministic. What he does not provide, however, are sufficiently good reasons for thinking that indeterminism in the moment of choice, in the face of certain motivational conflicts, will help secure claims of free agency. The compatibilist suspicion about the perils of trying to insert indeterminism into moments of volition, whether special and occasional SFAs or not, appears to be essentially correct. This kind of indeterminism looks likely to undermine agency; it does not help raise it to an esteemed form that we might claim to be free.

Chapter 3

Prospects for Compatibilism

Having seen the problems that arise in Kane's attempt to secure free will by positing special moments of indeterminism within the agent at special moments of choice, we now must turn to evaluate the prospects of compatibilist approaches to free agency. Here we are immediately confronted with the difficulty that compatibilism, in part because it has held sway as the mainstream position in the free will debates for so long, does not comprise just one clear position, beyond a commitment to the idea that our having free will is potentially compatible with the truth of some version of determinism. Despite this difficulty, my discussion and evaluation will attempt to extract meaningful themes and commonalities from amongst compatibilist approaches so as to generate a general view of their prospects.

In introducing the traditional problem or dilemma about free will and the traditional positions taken in the debate, I commented that compatibilism appears to be immediately confronted with a host of challenges, whereas libertarianism seems primarily confronted by one central challenge of making freedom of will 'compatible' with (some of) our actions being undetermined. Another way of interpreting this situation, however, is to say that the traditional problem/dilemma succeeds admirably in capturing the problems faced by libertarianism while, at best, it only captures one potential worry one might have about free will from a compatibilist perspective. The worry for compatibilists is that, under global determinism, actions appear as inevitable outcomes of events and states of affairs that predate the birth of the agent. This is a significant worry – indeed, my own reluctance to occupy a thoroughly compatibilist position stems in part from concerns about this way of looking at agency under conditions of global determinism – and we will return to it on more than one occasion in what follows.

Yet, beyond this particular (long distance) worry for compatibilists, it is evident that the traditional problem/dilemma represents a gross oversimplification (or misrepresentation) of what we can say about agency, action and freedom in a deterministic context. And it is, I will argue, largely because of these problems of oversimplification and/or misrepresentation that the challenges facing compatibilism become so readily apparent.

Agents owning actions

On a more local scale – the scale of an individual agent acting on a time scale of minutes, hours, days and even years – it is not at all obvious what should follow from the idea that, under determinism, "every action would be fixed by earlier events." (Lipton, 2004, p.89) For one thing, we have said nothing about the possible roles played by an agent in fixing one special class of events, namely the actions of that agent. Undetermined actions generate obvious problems for libertarianism in part because it is far from obvious how an agent could have played any role, let alone a significant role, in those actions coming about. But the idea that actions are fixed by earlier events seems to leave open innumerable possibilities for the agent to have played an important and ineliminable role in determining the occurrence of these actions.

This is especially clear when we consider two of the pivotal issues that I associate with concerns about free will – ownership and control. What I would call the primary compatibilist insight or conviction is the idea that claims of ownership of, and control over action are enhanced by (indeed may, *contra* libertarianism, even require) the agent playing a central role in determining that the action occurred. Not only is this an important insight into agency and action at what I have called a local scale, but it is equally important to acknowledge at long distance or a global scale. The idea that we make no difference to how things (including our actions) turn out because they are fixed by events in the past is not an idea that follows from the truth of determinism – it is just fatalism. While I suspect that many people have feared the implications of determinism because they have confused determinism with fatalism, we must be decisive in setting the latter thesis aside as largely irrelevant to our concerns. It simply does not follow from the (alleged) fact that my actions are determined by events in the past (determinism) that I play no significant role in those actions coming about (fatalism). The conceptual and theoretical space occupied by compatibilism is precisely one of spelling out the possible roles an agent might play in generating actions given the (or despite the possible) truth of determinism.

Returning, then, to the local level of explanation and justification of action, compatibilism appears to be on solid ground in claiming that the strongest kind of relation between an agent and their actions is one of determination. When asked to justify claims of ownership and control of, and responsibility for action, a compatibilist can claim that an ineliminable part of the factors determining the action includes (for example) the character and reasons of the

agent, and the agent's deciding to act out of, or on the basis of, that character and those reasons. This is evidently of special import in a context of contrastive explanation or justification, where an agent is tasked with explaining why they acted as they did instead of one or other alternative actions. The (compatibilist) agent can respond that they acted as they did, rather than in some other way, *because* they had this character and these reasons rather than some other character and/or reasons. The agent did not whimsically act on one reason rather than another, because the reasons they acted on were decisive in determining their action. (Of course, libertarian agent-causal accounts are also inclined to allude to an agent determining their actions, only in a context where the agent's character and reasons do not determine the outcome. Instead, the action is not determined by anything other than the agent. Exactly what this non-deterministic determining activity of the agent amounts to is for the most part unclear (if not mysterious) to both compatibilist and event-causal libertarians (such as Kane) alike.)

This relationship between agent, character, reasons and actions construed along compatibilist lines also aids in avoiding the particular problem of luck that confronts Kane (and, if in slightly modified form, other libertarians)[56]. In contrasting cases (such as in the 'cosmic re-runs' imagined in Chapter 1), a libertarian agent will act differently to how they in fact acted on at least some occasions, given exactly the same prior conditions, reasons, etc. One way of framing the problem of luck for the libertarian is that they do not seem to have access to any relevant difference between the cases that could explain the different outcomes – thus, whichever of the different outcomes (and their potentially different moral or ethical status) was the actual outcome seems, from this perspective, to be a matter of luck. For the compatibilist, by contrast, the (local) determination of action means that the outcome must always be the same, and different outcomes must be accompanied by relevant differences in the agent, their character, reasons and/or circumstances. At the local level, luck just doesn't come into the matter for compatibilism.

Negative Conceptions of Freedom

A further strength of the compatibilist position is that it encourages and grounds important distinctions between different ways in which agents might bring about their actions. This can

[56] For discussions of libertarian problems with luck, as well as some libertarian responses, see, for example, Double (1991), Strawson (1994), Clarke (1999, 2005), Haji (1999), Kane (1999a, 1999b) and Mele (1999, 2006).

be seen in the emphasis placed on what can be called negative conceptions of freedom – ways of understanding free actions by understanding what they are not. Many compatibilists emphasise the contrast between normal (free) action and action that is the product of coercion, manipulation or compulsion. That is, we should approach the question as to what acting of one's own free will consists in by, first and foremost, contrasting normal action with action emanating from an agent who is not free because they are subject to coercive, manipulative or compulsive influences. The agent with a gun held to their head lacks (compatibilist) freedom because they are not free to act on their reasons and character – the coercer has stacked the deck and forced on them reasons not of their making, with the consequences of acting in any way other than that prescribed by the coercer having an abnormal (and potentially deadly) import on the agent's deliberations. The agent whose psychology has been systematically manipulated without their knowledge of such manipulation does not act of their own free will, because their character and/or reasons are no longer entirely their own, but a product of the manipulation. The addict or obsessive compulsive acting under the influence of an irresistible desire does not act of their own free will because the desire will run its course to satisfaction no matter what the considered judgement or deliberations of the agent. Not only are these distinctions and contrasts significant and robust – they are important to any adequate account of free agency, including libertarian accounts.

It has even been argued that various compatibilist distinctions can survive an otherwise sceptical, illusion-based account of free will. Saul Smilansky (2000), in his book *Free Will and Illusion*, thinks that a core conception of 'up-to-usness' lies at the heart of our thinking about free will, and this core conception has both compatibilist and libertarian or 'ultimate' elements to it. While Smilansky (2000) argues that the libertarian or ultimate elements of the core conception cannot be defended, he thinks that compatibilist distinctions and categories remain important (and, thus, he rejects standard versions of hard determinism):

> *Not* denying all of the hard determinist case, I can still say that in order to be e.g. just we have to be partial compatibilists, and function on the compatibilist level, using compatibilist categories… It is morally crucial whether I had an opportunity to conform to the requirements of morality based on my autonomous reflections and local control, or not. These are, on one level, non-arbitrary moral distinctions, and any adequate moral order which respects me *as a person* will take them into account intrinsically, under a desert-based view. Even if I understand that there is no libertarian free will, I wish to live under a social order that requires that people be judged according to the paradigm of free will and responsibility… (Smilansky, 2000, p.92, italics in original)

For Smilansky (as for the hard determinist), the absence of libertarian free will means that no human agent is ultimately responsible for their lot in life. He thinks, however, that hard determinists go too far in their rejection of compatibilist conceptions of freedom and responsibility. From a perspective concerned with justice and just arrangements for society, Smilansky argues that we are better off living in a 'Community of Responsibility' where agents are given credit for their (morally) good efforts and acts, and where conditions that undermine compatibilist freedom and control are acknowledged as providing excuses from responsibility. Indeed, to fail to make these distinctions would, according to Smilansky (2000), only *add* to the injustice associated with the absence of ultimate (libertarian) responsibility.

While I have avoided the details of Smilansky's account, and while he clearly thinks that compatibilism cannot achieve all that it would hope as an account and defence of our having free will (an issue to which we will return shortly), his defence of compatibilist distinctions based on considerations of autonomy and local control is significant as a recommendation of the importance and robustness of these distinctions. Moreover, his arguments suggest that (negative) compatibilist conceptions of freedom should hold weight with both libertarians and hard determinists, suggesting at least one oasis of potential agreement in the entire free will debate.

Alternative Possibilities

Where compatibilism has traditionally fared less well is on the third pivotal issue in the free will debate – alternative possibilities. Indeed, it is probably fair to say that the central issue of contention between compatibilists and their incompatibilist counterparts (libertarian and hard determinist alike) has been the importance and interpretation of alternative possibilities as a condition for human agents having free will. I cannot hope to even briefly survey the outlines of this particular debate. Nevertheless, in looking optimistically at the prospects for compatibilism, I will highlight two trends in compatibilist approaches to alternative possibilities – (i) the intuitive plausibility of conditional interpretations of the requirement; and (ii) critical denials of the importance of alternative possibilities.

Perhaps the most widely endorsed compatibilist approach to questions about alternative possibilities is one that offers a conditional reading of the claim that an agent 'could have done otherwise' when acting freely. In essence, the argumentative strategy here is to insist

that the most plausible way to read a 'could have done otherwise' claim is as a conditional (subjunctive or counterfactual) claim about what an agent *could* have done had they willed differently. Thus, to say that I could have gone to the beach instead of staying here to work means both (a) that I was quite capable of choosing and going to the beach – I was physically capable, there were no obvious physical or psychological (coercive, manipulative, compulsive) obstacles or impediments to my choosing to go to the beach and acting on that choice; and (b) that I could have gone to the beach had I willed (deliberated/ reasoned/ desired/ valued/ etc.) differently. My will was free because I could have willed otherwise under relevantly similar circumstances. Counterfactually, I would (or might) not have stayed here to work instead of going to the beach if I had willed (deliberated, etc.) differently.

Incompatibilists have been notoriously unsatisfied with this interpretation of the alternative possibilities requirement, insisting that real freedom of will requires that more than one possible future of choice and action be genuinely (really, metaphysically) open to an agent at the time of choice and action. If, under determinism, there is and only ever has been one (genuinely, really, metaphysically) possible future for the agent's choice, then there are no *real* alternative possibilities and there is no *real* freedom.

The compatibilist, however, has two important responses for the dissatisfied incompatibilist. First, the compatibilist interpretation of 'could have done otherwise' has the virtue of emphasising considerations and distinctions that are of importance to compatibilist and incompatibilist alike, at least in so far as both sides share a commitment to a thesis of mental causation. For, if the relationship between agent and actions is to be interpreted along causal lines, then both sides will have to acknowledge the importance of the right kinds of patterns of counterfactual dependency in establishing and sustaining these causal claims (at least, on any account of causation that emphasises counterfactuals). Once again, it would seem that everyone (other than the epiphenomenalist) is likely to have to take on board significant portions of the compatibilist view of choice and action irrespective of their views on free will, thus leaving the compatibilist to ask why anyone would want more.

Second, the compatibilist can follow up on these more general observations by arguing in favour of the conditional interpretation as a means to securing the right kind of ownership and control of action necessary to sustain claims of agency, and thus free agency. As we have seen, libertarians have had great difficulty in sustaining the claim that ownership and

(especially) control can survive the insertion of indeterminism into the origins of an agent's actions. By focussing in on alternative possibilities, the compatibilist can try to show what is wrong-headed about libertarianism, largely independent of the problems with indeterminism, by exposing a crucial mistake in adopting an unconditional reading of 'could have done otherwise'. The mistake is something like the following: by adopting an unconditional reading of 'could have done otherwise', requiring that alternative possibilities be genuinely (really, metaphysically) available to the agent under *precisely* the same circumstances of choice, we can only be *weakening* the sense in which the agent's actions flow from (and connect with) their character, concerns and deliberations. After all, if the agent could have done otherwise under *exactly* the same circumstances, where these include their character (all that they brought to the situation by way of their psychological make up) and their deliberations (the actual path of reasoning they followed), it is far from clear that either their character or their deliberations played the kind of decisive role[57] in influencing their subsequent action that we would want if the agent is to claim ownership and control of the action.[58] Or rather, as we saw with the worries about indeterminism, it would seem that to the extent that we can allow a different outcome given the same precursors, we would be inclined to discount the level of ownership and control the agent might claim for their actions.

A key positive or constructive insight offered by the compatibilist about freedom and alternative possibilities is that we have the latter, in part, because of the kind of creatures that we are – unlike amoebae, *Sphex*[59], or a thermostat, we typically have an array of behavioural possibilities open to us in a given situation, and are sensitive to any number of contingencies or variables that may be relevant to the choice of action – and because we are able (under at least some conditions) to deliberately consider and evaluate these possibilities in the light of the contingencies so as to make an informed choice. Our having alternative possibilities in choice and action is, in this sense, metaphysically real to the extent that it tracks our genuine behavioural and cognitive capacities – the things we really can or could have done in a

[57] Agent-causal theorists foresee a decisive role for the agent in fixing their will; but, again, given the same character (in my inclusive sense) and deliberations, it is not clear to me what this could amount to.
[58] In Chapter 2, we saw that Kane (at least) can bring on board these sensible claims from the compatibilist approach, so long as the decisive willing of any given action has, in its history, some requisite number of SFAs relevant to the given choice situation. In my assessment of Kane, it emerged that even this strategy could not escape the problem, in that genuine SFAs and pseudo-SFAs (SFA-like choices involving a decisive judgement) seem only trivially different.
[59] The example of the wasp, *Sphex*, comes from Dennett (1984).

situation. Perhaps chief amongst these capacities is our capacity to generate or otherwise recognise options, and our capacity to make our actions track or respond to our reasons.

On the other hand, our *sense* of having alternative possibilities is also, in part, a function of the epistemic circumstances of deliberation and choice in so far as we do not know, prior to following through the course of our deliberations, which option from the array of possible actions we will settle on. For the compatibilist, the possible truth of determinism does not hold any threat to either of these bases for claims about alternative possibilities. It is not immediately obvious, at least *prima facie*, how or why our relevant behavioural and cognitive capacities would change if they turned out to operate on deterministic lines (or in a deterministic universe[60]); and, in the absence of any input to our deliberations based on deterministic predictions[61], our epistemic situation also remains apparently untouched by the truth or falsity of determinism.

For a range of compatibilists, then, our intuitions about alternative possibilities and 'could have done otherwise' can find a place in a compatibilist account of freedom as part of a story about the decisive, reliable and non-arbitrary connection between the agent's will and their actions. We claim ownership and control of our choices and actions because of the way in which these connect with and flow from our reasons, and the patterns of subjunctive conditionals and counterfactuals that are thus supported (as per a conditional reading of 'could have done otherwise') are an integral part of these connections between our reasons and our actions. If some other intuitions about alternative possibilities do not find a place in this scheme, it is those intuitions rather than our understanding of human agency that will, on these views, be at fault.

Living without alternative possibilities

The second compatibilist trend in dealing with alternative possibilities we can highlight takes this logic of faulting or challenging our intuitions to its logical extreme by critically interrogating the idea that we require alternative possibilities *at all* in attributing free will (and responsibility) to agents. On the one hand, Dennett (1987) famously uses (and reuses – see Dennett, 2003) the example of Luther on the church steps, proclaiming "Here I stand, I

[60] Although, as I will argue in Chapter 4, operating in deterministic universe turns out to be a significant problem for the compatibilist.

[61] That is, if we ignore the complications that might arise from having to consider the impact of knowledge of deterministically sound predictions of our own future behaviour into deliberations about that behaviour.

can do no other", as an alleged case in which the agent-declared absence of alternative possibilities makes no difference to our assessment of Luther as either a free agent or as a morally responsible agent. While interpretations of Dennett's example differ, Dennett's intended point seems quite straightforward: if we take Luther at his word, and we agree that Luther acted as a free and morally responsible agent when he nailed his declaration to the church door, then having alternative possibilities cannot be a necessary condition for free and morally responsible agency.

While 'occasionalist' libertarians like Kane[62] can claim to be in qualified agreement with Dennett, both agent-causal libertarian and hard determinist critics of compatibilism seem to have a case to answer here. In both real life and the hypothetical worlds of moral dilemmas, it seems we can think of any number of 'choice' situations in which only one possible path of action lies open to us, and this phenomenon needs to be squared with demands for the pervasive availability of genuinely (really, metaphysically) alternative possibilities as a necessary condition for free agency[63].

It could, of course, be argued that Dennett's point does not really intersect with the usual debate between compatibilists and incompatibilists over 'could have done otherwise'. Specifically, it might be argued that Luther's not being able to do otherwise is potentially at odds with the standard compatibilist reading of such a claim[64]. Presumably, on a conditional compatibilist reading, Luther could have done otherwise if only he had believed differently or had different values and concerns, and because he had enough by way of (negative) compatibilist freedoms to have acted differently if that had been his will. From this perspective, Dennett's insistence that Luther could not have done otherwise might be thought to detract from the insight pushed by advocates of a conditional reading, namely that even in Luther's case, we can find a sensible interpretation of the idea that he was free, in part, because he could have done otherwise.

However, I am inclined to interpret Dennett as being both sympathetic to the conditional reading of 'could have done otherwise' while also wanting to insist that we are not univocal

[62] See my Chapter 2.
[63] Dennett's (2003) preferred self-reflexive example from the domain of moral dilemmas involves his insistence that, given the chance to get a thousand dollars if only he would torture some people, he would (freely and responsibly) have no alternative but to refuse the money.
[64] I am grateful to Mark Leon for emphasising this point.

in our concerns about alternative possibilities. By focussing on cases like Luther's, Dennett is reminding us that outside of the philosophical debate over alternative possibilities, there are ordinary contexts in which an agent might find the insistence on being able to do otherwise puzzling, especially (though not necessarily exclusively) in cases of momentous decisions. "You insist that I must have been able to do otherwise; but I insist that if you were to place me in the same situation, or even a relevantly similar situation, I would do the same thing all over again!" Rather than undercutting the compatibilist emphasis on conditional alternative possibilities, this interpretation of Dennett suggests that both perspectives on alternative possibilities are grounded in the same (compatibilist) understanding of the conditions of ownership and control – namely that claims to *agency* are at their strongest when the force of the agent's will is most decisive. Thus: "I could have done otherwise had I thought or felt differently; but in this case, for me to have thought or felt differently would be for me to be an unrecognisably different person!" Under these circumstances, I (the agent) cannot see *me* 'surviving' as the agent and author of my action if the action were different.

While this position depends on a link to important and difficult questions about personal identity, it is not unusual for questions about identity to be raised in contexts where a philosophical theory suggests or implies a weaker connection between the agent and their projects, concerns and values than seems plausible to sustain. For example, one of the more effective critiques of utilitarian ethics offered by Bernard Williams (e.g. Williams, 1976) is that utilitarian calculations of the greatest good for the greatest number may well leave an agent in a position where the ethically prescribed response to a situation (based on the utility calculation) is so at odds with the character and projects of the agent as to leave their identity intolerably compromised. Similarly, I take it that Dennett is using Luther as a case in which the character of the agent could not, in a significant sense, survive even a conditional reading of the requirement of alternative possibilities (even if, as indicated above, we can allow that Luther has the requisite negative freedoms and capacity to act differently had his reasons been different). So, again, the idea is that the conditional interpretation of alternative possibilities is important but should not be pushed too far. If Dennett is right, an insistence on the pervasive availability of alternative possibilities is an incompatibilist hankering for their favoured mark of free agency that does not hold up to critical scrutiny; and it is a strength of compatibilism that it can capture what is plausible about the incompatibilist view (via the conditional interpretation of 'could have done otherwise') while avoiding its excesses of

strength (the unconditional or categorical reading) and scope (it is not a necessary feature of free and responsible choices).

On the other hand, a far more controversial challenge to the importance of alternative possibilities is due to Harry Frankfurt (1969), and plays a significant role in motivating the so-called semicompatibilism advocated by John Martin Fischer (1994, 2002; Fischer & Ravizza, 1998)[65]. The challenge is built on the construction of (appropriately named) Frankfurt-type examples whose purpose is to challenge the intuition that the availability of alternative possibilities (however construed) is necessary for moral responsibility[66].

The basic structure of a Frankfurt-type example involves an agent who is, in some way, at the mercy of another agency (a Cartesian evil demon, a 'benevolent' demon, an interfering advanced neuroscientist) that has the ability to intervene in the choices of the agent. So, in the case of Nell the Nasty Neuroscientist[67], we are asked to imagine that she has in some way mastered the ability to both monitor and intervene in the deliberations and choices of some unsuspecting human agent (whom I will call Ned). The neuroscientist is thus able to observe and track the course of Ned's deliberations in any given choice situation and, should she observe Ned tending towards a choice that is not to her taste, the neuroscientist can intervene and determine Ned's choice out of the options available as she sees fit. Given this setting, we have what looks like a clear case of (at least potential) manipulation that compatibilists and incompatibilists alike would recognise as a threat to free agency and responsibility.

The critical part of Frankfurt's thought experiment, however, lies in considering the cases where the nasty neuroscientist does not intervene – indeed, it lies in considering the cases where Ned happens to make choices that are to the liking of the neuroscientist on each and every occasion of choice during the time that he is under the (potential) influence of Nell. Frankfurt's contention is that, in these cases of non-intervention, agents like Ned are

---

[65] The key idea behind Fischer's position, and what makes it only *semi*compatibilist, is that he argues for a compatibilist position regarding determinism and moral responsibility while taking an incompatibilist stance (grounded in the Consequence Argument) on determinism and freedom. In short, Fischer thinks freedom requires alternative possibilities (which the Consequence Argument shows us not to have), whereas he finds Frankfurt-type examples persuasive in showing that we don't need alternative possibilities in order to have moral responsibility.

[66] Frankfurt introduced this type of example in Frankfurt (1969). For a variety of perspectives on the significance of Frankfurt-type examples, see the contributions by Fisher (2002), Ekstrom (2002) and Widerker (2002) that comprise a dedicated section of Kane (2002c).

[67] In Dennett (1984), he calls one of his bogeymen the 'Nefarious Neuroscientist'. I follow his alliterative lead, but not his choice of name.

determining their will – after all, there is no intervention – in a manner that is no different to how they would choose in the absence of the non-intervening controller Nell, while clearly lacking any categorical *or* conditional alternative possibilities for choice – should Ned have swayed towards any other option, Nell would have intervened to make things turn out as they in fact did. Thus, Frankfurt asks us to share the intuition that Ned is responsible for his choices, since these in fact turned out exactly as they would have in the absence of the neuroscientist, and despite the fact that he apparently lacked alternative possibilities.

Frankfurt-type examples arguably represent the strongest available challenge to the idea that alternative possibilities are important to free and responsible agency. Given that these thought experiments challenge both incompatibilist and compatibilist accounts of alternative possibilities, it is not surprising that the interpretations and responses provoked by the examples vary considerably. For some compatibilists, a possible line of response is to take onboard the basic interpretation offered above, and hold that a *truly* compatibilist account of free agency and responsibility just does not need to accommodate concerns about (categorical or conditional) alternative possibilities (other than by explaining that we have a *sense* of having alternative possibilities while we deliberate because of our epistemically limited position with regards to how our deliberations will run).

A different response would be to accept Frankfurt's inferences in so far as they pertain to questions about *responsibility* while being more cautious, or sceptical, about the implications for freedom. That is, one could accept Frankfurt-type examples as severing the tie between alternative possibilities and responsibility while still worrying (or insisting) that agents like Ned lack freedom in some sense important to our understanding of free agency.

Fischer's (1994; Fischer & Ravizza, 1998) semicompatibilism amounts to something like this second position, although Frankfurt-type examples are not the fundamental motivation for his position. Fischer is, for the most part, persuaded that the Consequence Argument (and similar arguments developed and discussed at length in Fischer, 1994) is successful in undermining our ordinary concept of freedom[68] in a deterministic universe because it shows that we lack regulative control. Frankfurt-type examples, however, suggest that we can rescue a

[68] More carefully, Fischer (1994) thinks we have two notions of freedom – freedom to do otherwise, and acting freely – and that incompatibilist arguments are successful in undermining the first notion given either determinism or the existence of God.

substantive notion of moral responsibility, based on a notion of guidance control, and this does not require the availability of alternative possibilities. The position is semicompatibilist both because it takes so much onboard from incompatibilist arguments about freedom, and because the position offers a compatibilism about responsibility while being incompatibilist in what it has to say about freedom[69].

Alternatively, one could accept the division between (at least some) questions of responsibility and questions of freedom while still maintaining (i) that these questions typically run together in non-Frankfurt-type cases, and (ii) that Ned lacks (compatibilist) freedom of an important kind. A compatibilist who favours a conditional reading of 'could have done otherwise' might insist that considerations of conditional alternative possibilities, by their very nature, involve hypothetical scenarios in which the agent willed differently (because of hypothetical differences in character, reasons, and/or circumstance). Given that issues of alternative possibilities are always hypothetical in this way, and given the interpretation that Ned lacks these alternative possibilities because of the pending interventions of the neuroscientist in all instances where Ned sways towards a different choice, Ned is not a (compatibilist) free agent. What the examples demonstrate is that compatibilist understandings of *responsibility* are sufficiently robust as to survive the agent's compromised freedom given the (contingent) lack of intervention in these imagined cases.

I am inclined to favour a version of this last position, primarily because it tries to secure what is sensible about conditional readings of 'could have done otherwise' when it comes to understanding claims of free agency while (as with Dennett's examples) qualifying the sense in which alternative possibilities are important within a compatibilist understanding of agency more generally. That is, given the primacy of issues of ownership and control in grounding agency, freedom and responsibility, it is perhaps not surprising that we can come up with (otherworldly) examples where contingently uncompromised ownership and control encourage intuitions of responsibility, while hypothetical considerations of alternative possibilities reveal possibilities for compromised ownership and control that would undermine agency and freedom. In terms of the current discussion, however, what is most pertinent is that if such a compatibilist response to Frankfurt-type examples can be sustained,

[69] See also my earlier note on Fischer's semicompatibilism. I find strong similarities between Smilansky's (2000) position and the overall gist of Fischer's work, in so far as Smilansky takes an incompatibilist line on freedom while also arguing that various compatibilist distinctions relating to just desert and worth remain important and justifiable.

then these thought experiments can safely be added to the list of considerations warning against an incompatibilist, categorical reading of 'could have done otherwise'.

A More Positive Conception of Freedom

Much of the compatibilist approach that has been surveyed so far has been of a negative nature – denials of or responses to various problems; understandings of freedom built in contrast to factors that negate freedom; critiques of the importance of alternative possibilities; and so on. When it comes to more positive, fleshed out accounts of free agency, it is that much more difficult to find common ground amongst compatibilists because these positive accounts tend to exhibit greater differences of opinion and detail. Nevertheless, in an attempt to evaluate the overall prospects for compatibilism, it would be useful to extract one or more common theme from these positive accounts that best captures the flavour of the picture of agency favoured by compatibilists. Just such a central theme can be found in the idea that free agency is marked, in various ways, by a capacity for *reflective endorsement*.

The idea that a capacity for reflective endorsement is important to human agency is hardly unique to compatibilism[70]. What distinguishes the compatibilist approach is the idea that some form of reflective endorsement, together with the negative conception of freedom discussed earlier (freedom from constraints of coercion, manipulation and compulsion), gives us enough by way of a conception of agency to save most (if not all) of what is worthwhile in our ordinary conception of free agency, irrespective of the truth of determinism. In one or other way, the agent who is both free of compatibilist constraints *and* able to reflectively endorse the choices that issue from their will *will be a free agent*. My task, at this point, is to put some flesh onto this skeletal theme of reflective endorsement, while allowing for differences of principle and detail that persist amongst competing compatibilist accounts[71].

[70] See, for example, Korsgaard's (1996) neo-Kantian account of autonomy and normativity cited in Chapter 9 below.

[71] One important difference amongst compatibilist approaches that I will tend to gloss over is that between more strictly structural accounts (such as Frankfurt's) and accounts with a stronger historical-genetic component (what Mele (2005) calls a history-sensitive compatibilism). On a strictly (or at least strongly) structural reading, an account like Frankfurt's does not have or need a historical or developmental story to tell about an agent and their motivational hierarchies – a suitable alignment between higher and lower order volitions will suffice for a will to be free (but see my comments on identification in Frankfurt's account, below). Dennett also tends to favour a view that does not require historical conditions to be satisfied, even if such conditions usually are satisfied in ordinary (and especially non-thought-experiment) cases. See, for example, exchanges between Dennett and Mele in Dennett (2003, 2005) and Mele (2005).

I take the following quote from Dilman (1999) to be expressing an important idea about human agency that, barring various possible disputes over details, underpins much compatibilist thinking about free agency:

> Even if at first I simply accept, indeed swallow, my parent's (sic) values and precepts, what I learn gradually enables me to reflect on them, come to understand their significance in relation to the situations with which life confronts me. While I weigh those situations in terms of these values, at the same time those situations weigh those values for me in the light of other forms of significance I pick up in the course of my upbringing and other contacts. It is in this process that I begin to come to myself, come to have a self I can come to, and at the same time come to own the values or reject them in favour of others I come in contact with and make my own… I am no longer a product of my upbringing. In the course of it I acquire a mind of my own and a will that is mine… I acquire the capacity to weigh and accept or reject what I am given… [M]ore and more I participate in this process of learning and growing up, that is in my own formation, as I acquire independence of thought and come to myself. I thus come to own myself, and it is as myself, as an individual, that I think, consider, take decisions and act… (Dilman, 1999, p.246)

While this view of agency need not be read as inherently compatibilist, I will unpack it in a manner that highlights what compatibilists might find most salient and attractive about it.

In its essential details, the picture presented is that of an agent growing into themselves, gradually taking ownership and control of their "values and precepts" through processes of reflective endorsement. For a typical compatibilist, there is no magic moment at which a free human agent emerges from the complex of developmental processes that have constituted their upbringing; nor are there magic moments of indeterminism (be these SFAs or instances of agent-causation) that gradually or instantly mark the emergence of a free agent. Instead, there is (just) a gradual process through which distinctively human agency emerges, enabled by capacities for reflection and learning, such that the values and precepts imbibed during development come under reflective scrutiny in the light of situations and circumstances that are encountered, including encounters with different systems of value and belief that the agent finds in others. While the agent is always coming to a situation from a point of view structured and shaped by their current values and precepts, the reflective agent can also assess that point of view through comparisons with other "forms of significance" and systems of values and precepts that are encountered and/or encouraged by life experiences and other people, weighing these against each other, and coming to own or reject and replace the values and precepts that were the agent's starting point. It is through this gradual, temporally extended process of reflective engagement with the world and others, accompanied by various endorsements and rejections/alterations of the values and precepts brought into these situations, that the agent comes to own the person that they are with the (continually

evolving) character that they have. And in acquiring ownership of themselves, they gradually take ownership of their will and the actions that flow from their will.

In keeping with the compatibilist theme of focussing on the normal case, I take it that various important compatibilist approaches to free will and agency presuppose something like this view of reflective endorsement as a background condition to claims that particular instances of willing and acting constitute choices made by free and responsible agents. This theme in compatibilist thinking can be traced back at least as far as Locke:

> For, the mind having in most cases, as is evident in experience, a power to *suspend* the execution and satisfaction of any of its desires; and so all, one after another, is at liberty to consider the object of them, examine them on all sides, and weigh them with others. In this lies the liberty man (sic) has; and from the not using of it right comes all that variety of mistakes, errors, and faults which we can run into in the conduct of our lives, and our endeavours after happiness; whilst we precipitate the determination of our wills, and engage too soon, before due examination… we have the opportunity to examine, view, and judge the good or evil of what we are going to do; and when, upon due examination, we have judged, we have done our duty, all that we can do, or ought to do, in pursuit of our happiness; and it is not a fault, but a perfection of our nature, to desire, will, and act according to the last result of a fair examination. (Locke, 1690/1975, pp263-264 - Section 47)

The space of reflection and evaluation of our motives and desires is thus, for Locke, the space in which we should locate distinctively human freedom. It is a freedom from wantonness – from agency where desires merely compete, on the basis of strength and persistence, for control of action. Human agency is marked by action on the basis of motives that have survived reflective examination and have been endorsed by judgements of reason.

More contemporary forms of compatibilism have carried this theme forward in a variety of guises[72]. For Frankfurt (1971), an agent's will is free when the agent has an appropriate second (or *n*-) order volition that a certain first- (or lower-) order desire be effective – they want that their will should be constituted by the given lower-order desire. While this should not be mistaken for the claim that such higher-order endorsement takes place occurently in or at the moment of willing, reflective endorsement of *some kind* within the agent's psychological make up is being posited as the grounds for having a will that is free: 'reflective' because the higher-order volition takes a lower-order desire as its object, and 'endorsement' because the volition is constituted by the desire that the lower-order desire be effective. Moreover, given the well-known problems associated with hierarchical accounts[73], and the vagueness of notions of identification employed by Frankfurt, it is not unreasonable

[72] For a useful discussion, see Haji (2002).
[73] See Haji (2002).

to suggest that the picture sketched by Dilman (1999) is the kind of image of reflective agency that needs to be in place if hierarchical accounts are to be suitably fleshed out.

Gary Watson, at least in his most well-known discussion of his views (Watson, 1975), arguably subscribes to a similar view of the grounds of free will, even while he explicitly contrasts his account with the hierarchical position of Frankfurt (1971). Watson (1975) claims that agents have a valuation system and a motivational system, and that the possibility of unfree action arises, not because of the possible truth of determinism, but because of the possibility of an agent's motivational system being effective in producing action independently of the judgements issuing from the valuation system – that is, agents are characteristically unfree when their desires are effective independent of their considered or better judgements. Watson views the link between valuation and free agency as follows:

> *The valuation system* of an agent is that set of considerations which, when combined with his (sic) factual beliefs (and probability estimates), yields judgements of the form: the thing for me to do in these circumstances, all things considered, is *a*. To ascribe free agency to a being presupposes it to be a being that makes judgements of this sort. To be this sort of being, one must assign values to alternative states of affairs, that is, rank them in terms of worth. (Watson, 1975, p.105, italics in original)

Thus, while being in possession of a valuation system is not all that there is to being a free agent, Watson (1975) clearly thinks that the reflective deliberation and endorsement characteristic of having such a system is a significant (positive) mark of what it means to be a free agent. Moreover, given that an obvious line of critique for this view would be one that questions the origins and status of the *inputs* to the valuation system, it is again plausible to suggest that Watson would have in mind something like the picture sketched by Dilman (1999) as the relevant, normal background conditions from which an effective valuation system of a free human agent emerges. Indeed, Dilman's description of the gradual encounters of an individual with competing systems of "values and precepts" provides a (skeletal but) plausible developmental account of how the valuation system of a human agent might emerge and, to some extent, begin to diverge from the motivation system, the latter presumably having been structured by biology, early experience and the internalised value systems of the parents.

Dennett's (1984, 2003) brand of compatibilism, while increasingly couched in terms of an evolutionary account of human agency and morality, also emphasises reflection and self monitoring as a hallmark of free agents – as he puts it, "the transition from oblivious agents to minded, reflective agents" (Dennett, 2003, p.261). The space of distinctively reflective

human agency is the space of giving and demanding reasons for action, both to and of ourselves and others. Free agency is thus tied to the challenges of regulating and coordinating our behaviour and that of others with whom we interact, and it is our capacity for reflection and reason-giving that lies at the heart of such efforts:

> Our evolved capacity to reflect gives us – and only us – both the opportunity and the competence to evaluate the ends [we might pursue], not just the means. We have to use our current values as the starting point for any contemplated revaluation of values, but from our perspective on our current hilltop, we can formulate, criticize, revise, and – if we are lucky – mutually endorse a set of design principles for living in society… [W]e may be able to discover some adjustments in our current design that have some hope of carrying us to higher summits. And unlike all other species [for whom only 'blind' natural selection could tackle such problems], these are problems *for us*. We actually work on them, devoting time and energy to them. We gather information relevant to them, explore variations on them, and debate their merits knowing that our reflections will actually help determine which trajectory our future holds. (Dennett, 2003, p.268, italics in original)

And the starting point for such reflective attempts at self and social design is roughly the same as the one described by Dillman:

> A proper human self is largely the unwitting creation of an interpersonal design process in which we encourage small children to become communicators and, in particular, to join our practice of asking for and giving reasons, and then reasoning about what to do and why. (Dennett, 2003, p.273)

Once again, we find a sketch of the gradual, progressive emergence of the adult human agent, able not only to contemplate alternative futures (ends) and various means to realising those futures, but also to evaluate, revaluate and endorse courses of action through the practices of reflection and reason giving.

Themes of reflective endorsement can, unsurprisingly, be found in accounts of agency and freedom that explicitly emphasise autonomy as their critical issue. Mele (1995), while officially agnostic about the truth of compatibilism[74], spends a significant amount of time developing a compatibilist account of autonomous agency. Mele's (1995) basic argument in *Autonomous Agents* is that an appropriate combination of self control and autonomy is what is required for an agent to qualify as a free agent[75]. The importance of self control in this picture of agency is closely tied to the issue of action in accordance with an agent's decisive best judgement of what to do:

> [S]elf-controlled individuals are agents possessed both of significant motivation to conduct themselves as they judge best and of a robust capacity to do what it takes so to conduct themselves in the face of (actual or anticipated) competing motivation. (Mele, 1995, p.5)

[74] In Mele (1995) and Mele (2006), he offers what he thinks are the most reasonable compatibilist and libertarian positions with regards to autonomous, free agency.
[75] Part I of Mele (1995) develops the argument that self control is not sufficient to give us autonomy; Part II then develops compatibilist and libertarian conditions for autonomy.

The salient contrast for self-controlled individuals are akratic or weak-willed individuals who:

> …suffer from a deficiency in one or both of these connections. Human beings *wholly* lacking self-control are at the mercy of whatever desires happen to be strongest, even when the desires clash with their better judgements. (Mele, 1995, p.5, italics in original)

Given this conception of self-control (and its converse in *akrasia*), one can read Mele's (1995) overall argument as claiming that at least one conception of agency plus reflective endorsement – that is, a conception on which reflective endorsement amounts to simply making decisive best judgements about what to do, and then acting in accord with these judgement – is in fact not sufficient to give us a robust free agency. Self-controlled agents must also be autonomous agents, capable of autonomous reasoning and judgement, if they are to have a robust and meaningful freedom.

Mele (1995) is at pains to point out that he does not suppose *all* free actions of a self-controlled agent need meet the conditions he offers for autonomous actions. The logic of his position is rather that if any actions qualify as free, these should include the class of 'full-blown' intentional actions that might satisfy his compatibilist conditions for autonomous action:

> Full-blown, deliberative, intentional action… involves some psychological basis for evaluative reasoning (e.g. values, desires, and beliefs); an evaluative judgement that is made on the basis of such reasoning and recommends a particular course of action; an intention formed or acquired on the basis of that judgement; and an action executing that intention. (Mele, 1995, p.177)

Since, amongst other concerns, both the psychological basis for reasoning and the deliberations and judgements themselves might be compromised in various (especially manipulative) ways, his account of psychological autonomy and autonomous action is intended to flesh out the conditions that need to be added to our picture of a self-controlled deliberator if the latter is to qualify as a free (autonomous) agent.

In Part II of *Autonomous Agents*, Mele (1995) develops and defends a set of conditions for autonomy that he thinks needs to be added to a realistic conception of a self-controlled human agent. I will focus on describing (not defending) the conditions that he offers for a compatibilist conception of autonomy (one additional condition, involving indeterminism, is added to yield a libertarian conception of autonomy). Mele (1995, pp.223-4) asks us to consider a budding agent, Betty, who is six years old and who is afraid to go into the basement of her house alone, especially if the basement lights are off. Betty is aware that nothing untoward has ever happened to her in the basement, and that her older sister does not

seem to have any fear of being in the basement. Betty "comes to view her fear as a 'babyish' one, and she decides to try to eliminate it" (Mele, 1995, p.223) by visiting the basement periodically until the fear disappears. In a later chapter, Mele (1995) summarises the compatibilist conditions he offers as sufficient for Betty to have freely tried to eliminate her fear:

> (1) She was not compelled to have any of the pro-attitudes that grounded her judgement that it would be best to try to eliminate her fear, nor were any of those attitudes coercively or self-oppressively produced; (2) she was not cut off from autonomous reasoning by her doxastic condition; (3) the reasoning that led to her decisive better judgement was reliable reasoning; and (4) the judgement issued unproblematically in a corresponding intention that issued smoothly in a corresponding intentional attempt to eliminate her fear. (Mele, 1995, p.247)

As I interpret these conditions, (1) captures much that is standard in compatibilist accounts of negative freedom, while (4) touches on issues relating to self-control and non-deviant relations between Betty's reasons and reasoning, and her actions. Conditions (2) and (3) point to relevant features of Betty's epistemic position (especially the information and beliefs upon which she deliberates), and her status as a reliable deliberator when it comes to means/end reasoning. For Mele (1995), conditions (1) to (3) provide a candidate set of sufficient conditions for psychological autonomy conceived along compatibilist lines, while (4) is an addition to provide sufficient (compatibilist) conditions[76] for autonomous action.

Reflective endorsement thus fits into a more complex picture of agency on autonomy-based theories such as Mele's. In a sense, the mere presence or occurrence of reflective endorsement (by way of decisive best judgements) is seen as being insufficient as a ground for free (autonomous) agency because the inputs to reflections and deliberation (condition 1), the epistemic circumstances of the agent (condition 2), and the deliberations themselves (condition 3) might be compromised in various ways that threaten freedom. Nevertheless, the additional complexities are ultimately in service of securing the autonomous status of the agent's reflective endorsements of their will and their actions by way of decisive best judgements of what to do, and to this extent (at least) reemphasise the centrality of this theme within compatibilist thinking about free agency.

Moreover, Mele (1995) characterises the emergence of autonomous agency in ways that resonate with the description of development offered by Dilman (1999). According to Mele (1995), children clearly acquire and develop ever greater degrees of control over their bodies

[76] Mele (1995) offers candidate sufficient conditions for both a compatibilist and an incompatibilist brand of autonomous action.

and minds, including increasing (if modest) competence in identifying means to ends (think of young Betty and her insight that forcing herself to go into her basement might make an effective means to the end of reducing her fear of the basement). For Mele, the key issue is not one of children creating themselves as agents *ex nihilo* or overnight, but of trial, error and practice:

> Our earliest limb movements are mere flailings, but the flailings themselves play a role in our gaining control over the motions of our limbs. Our earliest sequences of practical thoughts might best be viewed as mental flailings, flailings that play a role in our becoming competent practical reasoners. Eventually, we are able to represent options, to select among them, and to take steps toward the selected goal. By that time, we are beyond mere bodily and mental flailing, and we are able to *choose*. (Mele, 1995, p.228; italics in original)

Not only do we progressively acquire the capacity to choose, but our choices and actions themselves begin to shape who we are (and, to the extent that such choices and actions are intentional, some of this shaping is itself intentional). Citing Betty as an example, Mele (1995, p.229) suggests that such attempts at intentional self-modification could enhance self-esteem and impact on the view Betty has of just what it is about herself and her world that she might be able to control and/or change. (One might allow that even intentional action that is not explicitly self-modifying in its goals could have similar effects, at least to the extent that the impact of the action on the world and on the agent is noted by the agent themselves.)

In various different ways, as we have seen, a range of compatibilist accounts of free agency centred around notions of reflective endorsement provide a plausible and, in many ways, appealing sketch of the form of agency to which normal adult human beings might come given various background conditions in their development, and the emergence and exercise of various cognitive and behavioural capacities. In contrast to varieties of libertarianism, compatibilists do not see a need to posit special indeterministic moments of either character shaping or decision-making (as in Kane's SFAs, or in agent-causal accounts of decisions more generally) in order to secure the right kinds of ownership and control of action we might require of a free agent. And, while it is not altogether clear to what extent accounts of free agency should be normative or prescriptive, the compatibilist agent exercising their capacity for reflection, endorsement and self-modification is arguably also an ideal advocated in many different ways within various systems of (especially Western) thinking.

Lingering Concerns

Despite the obvious appeal and good sense behind much compatibilist thinking, both in terms of its emphasis on negative freedom and the variations on themes of reflective endorsement

as a distinctive and desirable feature of human agency, a number of lingering concerns should be noted that may qualify any wholehearted endorsement of compatibilism. The three concerns I have in mind involve (i) lingering questions about 'ultimate up-to-usness'[77]; (ii) concerns about 'creeping exculpation'[78]; and (iii) questions as to whether reflective endorsement gives us a rich enough, and empirically defensible account of human agency on which to ground claims about freedom.

It may sound strange, at least to a compatibilist, to raise questions about 'ultimate up-to-usness' as a source of lingering concern given the extent to which compatibilism is typified by a rejection of demands for ultimacy – whether these demands are framed in terms of ultimate responsibility, unmoved-mover ultimacy, regress-halting ultimacy, or other similar incompatibilist concepts. As we saw in his critique of Kane, Dennett (2003) is inclined to view the search for a regress-stopping form of ultimacy as being on a par with biologists searching desperately for a Prime Mammal lest, in the absence of such a creature, we should have to conclude that there are no mammals. Compatibilist accounts that emphasise negative freedom reject demands for ultimacy as misguided, if not incoherent, given the argument that freedom is more sensibly viewed as an absence of external constraint, coercion, manipulation, etc. And I take it that two of the central ideas motivating the theme of reflective endorsement in compatibilist thinking are that, first, the search for ultimacy construed along various incompatibilist lines of thinking is misguided and doomed to either failure or giving us something of questionable or negligible value; and, second, that we need not despair because the average agent's capacity for reflective endorsement means we can, as it were, progressively build up enough 'up-to-usness' over time to secure the kind of ownership and control we would want a free agent to have.

Nevertheless, I think that agnostics and newcomers to the debates on free agency may well be swayed by concerns about ultimate 'up-to-usness' for a number of interrelated reasons, most aptly captured by talking about luck[79]. Libertarians have what one might call short-term or local level problems with luck – given the posited role of indeterminism in (at least some) of our decisions and actions, it seems to many that it is problematically a matter of luck or chance how those decisions or actions turn out, thus compromising our ownership and control

[77] This expression, and the idea of an ultimate perspective, is discussed in Smilansky (2000).
[78] Dennett (1984, 2003) uses this phrase when discussing questions of escaping responsibility.
[79] See Mele (2006) for an extended discussion of questions of luck in relation to free will.

while trying to secure ‘real’ alternative possibilities. Compatibilists, typically, have a more long-term or global level problem of luck. If, with Smilansky (2000), one accepts that some idea of ‘ultimate up-to-usness’ is (*contra* compatibilism) a fundamental part our conception of free agency, then the absence of libertarian free will should persuade us that we lack this ‘ultimate up-to-usness’. From an *ultimate* perspective, it will still be the case that an agent’s having negative freedom (absence of coercion, manipulation, constraint, etc.), and their having and actually exercising capacities for reflective endorsement, will all boil down to questions about a lottery of factors involving their parents, their genetic endowment, the circumstances into which they are thrown, and (perhaps) years of critical development that will shape who they are and what they can become.

Note that this is not simply a restatement of the puzzles raised in the sceptical dilemma about free will, or in the Consequence Argument, arising from the idea that determinism makes all these things (and all decisions and actions) an inevitable consequence of the state of the universe preceding an agent’s birth. The latter obviously provides one kind of ultimate perspective on agency, and it can be used to generate long-distance problems of luck. Instead, I have tried to emphasise questions of luck in terms of an agent’s endowment (both initial and as a result of development) that they might bring to any opportunity for choice, or reflective endorsement, or self-shaping activity, etc.

Think, for instance, of Mele’s (1995) example of young Betty cited earlier. The kind of long-distance questions about luck that an in- or non-compatibilist could raise in her case might point to her *good fortune* in being a six-year old with sufficient capacity to reflect on her fear of the basement, and sufficient ability to reason her way through to a strategy that has a reasonable chance of helping to reduce that fear. In addition, she is *fortunate* to the extent that she merely has a phobia of unknown origin, as opposed to having been born into a family of abusive, manipulative parents who actively cultivate and maintain phobias in their children; or having an older sister who might deliberately deceive her about the irrationality of her fear of the basement (perhaps there is something to fear down there, and her older sister has only pretended to visit the basement so as to maliciously enjoy her younger sibling’s terror when she follows her ‘lead’). In short, Betty’s negative and positive ‘freedom’ could have been compromised in various ways as a function of luck and accidents of circumstance; and so the fact that her agency is not lacking these ‘freedoms’ appears, from a suitable distance, equally to be a matter of luck. If having and exercising these freedoms is a ultimately a matter of

luck, then whether or not we exercise the associated capacities seems a matter of chance, and (further) does not look like something deserving of praise or blame.

Compatibilists can resist this line of reasoning by taking a more actively *ahistorical* view of free agency. That is, compatibilists can opt to take a more synchronic view of free agency whereby an agent's freedom is to be located in the structure and operations of their will at a given point in time *plus* the absence of certain precursors to deliberation and choice (the negative freedoms). As noted earlier, Frankfurt's hierarchical account can be read in this way – it is the synchronic meshing of higher-order volitions with lower-order effective desires (plus negative freedom) that makes a choice free, not the long-term developmental history of the agent. But I think that any such move by compatibilists will tend to undermine much that is plausible in their attempts to replace untenable ultimate conceptions of agency with more emergent, developmental perspectives on agency and freedom. Problems of luck in the context of such emergent accounts need to be faced and, where they cannot be resolved, at least honestly acknowledged.

The second area of concern for compatibilism involves what Dennett (2003) calls the spectre of 'creeping exculpation' – worries about the potentially dwindling sphere of free and responsible agency in the face of ever widening arrays of excusing conditions. To the extent that the apparent availability of excusing conditions follows from our ever greater empirical knowledge of humans, their development and functioning, it might be thought that questions and worries about creeping exculpation need to be addressed by compatibilist and libertarian alike. If, for example, we find that our genetics really can stack the deck for some individuals in ways that make certain behaviours (or dispositions) practically inevitable for those agents, then all participants in the debate over freedom and responsibility would need to factor this into what is said about our practices of attributing responsibility and blame.

There are, however, at least two reasons for thinking that compatibilists have a little more to worry about here than their libertarian counterparts. First, libertarians can (in principle[80]) draw on whatever role they have given to indeterministic processes in order to posit some space for a break between the agent's dispositions and their subsequent choices and actions.

[80] That is, on the assumption that their chosen libertarian account is sustainable in the face of compatibilist and hard determinist criticism.

This won't matter for those things which turn out to be truly inevitable[81], but the libertarian (and agnostics) might want to argue that it will help in a range of other cases.

Consider the possibility that an advanced psychology (or interdisciplinary mind science, if you like) begins to offer us clear probabilities for certain behaviour in particular individuals – say, probabilities for domestic violence in men that fit a certain profile[82]. In an individual accused of domestic violence where these prior probabilities were high (greater than 0.5, say), a libertarian might make a case for responsibility (because the behaviour was undetermined prior to the agent's choosing it), whereas a compatibilist might have to acknowledge a set of excusing conditions because of the high prior probabilities[83].

The second reason for thinking compatibilism has a bit more to worry about with regards to creeping exculpation is because, on average, compatibilism is at greater risk than libertarianism of setting the bar too high when it comes to qualifying as a free and responsible agent. Whatever libertarianism's difficulties, it is easier for a libertarian to see most human agents as acting freely and responsibly most of the time because the relevant ingredients that make for a free agent need not be in short supply. Agent-causal accounts in general, and Searle's indeterministic account of free will, tend to see indeterminacy as a regular feature of human choice and action; and there seems to me little in Kane's account to think that SFAs are in short supply in the average human agent. Compatibilists, no doubt, would like to claim that their accounts are similarly easy for the average human agent to satisfy; and we have noted that there is something appealing and attractive about the sketch given by Dilman of how we come to ourselves as agents as we develop, grow and mature. The question is whether the compatibilist can make these claims plausible.

I will limit my discussion on this issue here, both because (as we will see in Chapter 4) I think the real problem with compatibilism lies elsewhere, and because (as I will describe in

---

[81] Dennett (2003, p157) mentions the example of Huntington's disease as a case of true inevitability.

[82] In this imagined scenario, the profile might be exclusively psychological/ psychosocial, or it might (as per the imagined interdisciplinary science of the mind) incorporate aspects of psychological, neuroanatomical and neurophysiological detail.

[83] Obviously, there is much more to be said on either side about such a case. Compatibilists are, in particular, unlikely to accept that they must accept the probabilities as excusing conditions; and they are also unlikely to concede that the libertarian has any advantage, given the apparent difficulties libertarians have with indeterminism. Moreover, there is much more that needs to be said about explanations involving probabilities and probabilistic causation. The only crucial point here is that, *prima facie*, the libertarian tends to posit a break or gap between the prior probabilities and the outcome, where the compatibilist must factor in the probabilities as is.

Chapter 5 and discuss in Chapter 9) there is, to my mind, a bigger issue here of risking portraying ourselves as hyper-rational, hyper-reflective agents. When we do this, we risk disqualifying ourselves from being candidate free agents; we also risk aspiring to normative ideals that might be unattainable and, in certain ways, undesirable. I will argue, later, compatibilism seems especially prone to running this risk. But let me say something quick on this issue in relation to creeping exculpation.

Consider Mele's (1995, 2006) approach to developing an account of autonomy and free agency. As we have seen, the notion of an ideally self-controlled agent plays in important role in his account. In *Autonomous Agents*, Mele (1995) specifically lays out a project of asking what needs to be added, in the case of an ideally self-controlled agent, in order to get an autonomous agent[84]. Mele thinks that most normal, healthy adult humans evidence most (if not all) of the features of the ideally self-controlled agent to varying degrees at least some of the time. He does not, as far as I can tell, claim that we are ideally self-controlled agents. Given our manifest deficiencies in self-control, that seems appropriate.

One consequence of this way of setting up his project is that it must remain unclear to what extent the account applies to real-world human agents. In relation to questions of creeping exculpation, what we are then faced with is the possibility that securing the integrity of our account of autonomy (Mele, 1995) or free agency (Mele, 2006) comes at the cost of disqualifying many, most, or just too many real agents in the world. We are, that is, running the risk of making available increasing numbers of excuses for agents who claim, through no fault of their own, to not meet the standards for freedom and responsibility on offer. I use Mele simply as an exemplar of how setting out compatibilist standards for free and responsible agency might quickly leave behind real-world agents that non-compatibilists might have thought reasonable candidates (as on, for example, less demanding libertarian accounts).

A committed compatibilist, convinced of the virtues of their account, has many potential responses at this point. At one level, the compatibilist project might be viewed as a conceptual project of demonstrating how there might be free agents in a deterministic universe, with no commitment (or care) whatsoever as to whether or not humans, or humans

[84] This is the goal of Part II of Mele (1995).

in general, are such agents. This is a perfectly reasonable view of things; it is just (presumably) not the reason why most of us were interested in questions about free will: we thought we had (or might have) it.

A different kind of compatibilist response would point out that, if fewer humans land up qualifying as free and responsible on a compatibilist account that is (or is mostly) correct, there is no point in complaining about it or trying to deny the fact. Indeed, the politics of treating as free and responsible those who lack the necessary and/or sufficient capacities and features for such agency tend to be conservative, intolerant and harsh.

This second response is also reasonable. Given a correct (or mostly correct) account such as that imagined above would require revision and accommodation on many fronts. In the absence of decisive evidence for any particular compatibilist account, however, not to mention lack of decisive arguments in favour of a compatibilist approach in general, it counts against a compatibilist project if it threatens to make the ordinary extraordinary. As indicated in my comments on the previous response, free will is a puzzle for many because we think we have it, but we're not sure what to say about how that comes to be.

My third lingering concern about compatibilism is the worry that varieties of compatibilism, grounded in notions of reflective endorsement and operating (for the most part) in the space of belief-desire psychology, reasons and reason-giving, do not offer a rich and empirically-detailed enough conception of agency to be able to address challenges to our conception of agency and freedom stemming from philosophy and, especially, empirical science. How, other than by way of empirically-informed discussions of development and psychological functioning, can we sensibly respond to long-distance worries about luck? How can we satisfactorily deal with claims of creeping exculpation grounded in genetics, or in particular psychological or socio-cultural theories, except by trying to deal with at least some of these on their own terms?

In short, it is unclear how we can deal with any of a range of worries and challenges about our potentially being passively moulded into agents, variably capable of (potentially limited) reflective endorsement, if we are not better able to address these issues in the light of empirically-informed ideas about development, action, agency, and a range of other issues that span the increasingly artificial divide between philosophy, on one hand, and psychology

and other interdisciplinary sciences of the mind-body on the other. Nor is it clear how we might respond to sceptical challenges whose evidential base is not conceptual and philosophical but unambiguously empirical – such as the sceptical arguments regarding conscious will, volition and agency that are presented in Chapter 6.

There is, however, a further and deeper concern about compatibilism. It is a concern that overlaps significantly with those mentioned above that relate to ultimate up-to-usness and long-distance problems of luck. In essence, it is a concern that arises from conducting what should be the ultimate test for compatibilism, namely assuming the truth of determinism, and imagining the implications of this for suitable samples of actual, possible and conceivable agents. One such test, recently proposed by Alfred Mele (2006) in the form of what he calls the zygote argument, presents a challenge to compatibilism that is not easily answered. Indeed, as I will argue at length in the following chapter, there is good reason to think that compatibilism cannot answer this argument in a way that would satisfy the as yet uncommitted, not to mention incompatibilist, parties to the debate over free will.

Chapter 4

Zygotes, Manipulators, and the Failure of Compatibilism

In *The View from Nowhere*, Thomas Nagel expresses the following frank sentiments about the traditional problem of free will:

> I change my mind about the problem of free will every time I think about it, and therefore cannot offer any view with even moderate confidence; but my present opinion is that nothing that might be a solution has yet been described. This is not a case where there are several possible candidate solutions and we don't know which is correct. It is a case where nothing believable has (to my knowledge) been proposed by anyone in the extensive public discussion of the subject. (Nagel, 1986, pp112-3)

In *Philosophical Explanations*, Robert Nozick prefaces his discussion of free will with a qualification that almost has the ring of an apology:

> Over the years I have spent more time thinking about the problem of free will – it felt like banging my head against it – than about any other philosophical topic except perhaps the foundations of ethics. Fresh ideas would come frequently, soon afterwards to curdle. (Nozick, 1981, pp293)

These are candid, somewhat frustrated, and arguably pessimistic sentiments being expressed by two eminent philosophers who have taken the time and effort to nevertheless include what they themselves seem to regard as less than satisfactory discussions of free will in each of their important books.

Of course, it could be argued that a thorough search might reveal equally frustrated and/or pessimistic sentiments being expressed by equally eminent philosophers about consciousness, or mental causation, or intentionality, to choose but a few examples. Philosophical puzzles are, well, puzzling. But there are hints within these quotations that there might be some deeper difficulty/ies that we encounter within the traditional debate over free will that make/s decisive movement or resolution within the contours of the debate seem impossible.

Nagel (1986) thought that nothing believable had yet been proposed by way of candidate solutions. Even allowing for the passage of more than twenty years, there is a strong likelihood that Nagel might assent to the same judgement about the debate today. This would no doubt be contested by participants on all sides of the debate, and perhaps most especially by the broad church of compatibilism, whose devotees might try to cash out claims to orthodoxy in terms of greater believability. And a libertarian like Kane might, in part, defend his position by reminding us both of background assumptions which tend to close us off to the plausibility of a Kane-type account (most especially, our tendency to equate

‘indeterministically caused’ with ‘uncaused’), as well as of the empirically open hypothesis built into his account that awaits future testing and evaluation.

The burden of the argument in Chapter 2 was to suggest that Kane’s account provided libertarians with their most constructive and least mysterious attempt yet to insert indeterminacy into volitional processes in such a way as to make good the claim that this was both necessary and desirable in securing claims of freedom of the will. But, on my evaluation, Kane’s attempt fails. Its failure is, I think, instructive because it reaffirms the basic idea that inserting indeterminism into volition itself (relative to a background of largely or potentially deterministic causation) can’t secure freedom whilst it at the same time undermines (or at least weakens) our claims of agency, ownership and control. From this perspective, libertarian attempts to secure freedom by adding indeterminism into moments of volition seem fundamentally misguided, despite the many attempts (including Kane’s) to convince us otherwise.

The discussions of the previous chapter suggested that, on this and various other fronts, compatibilism tends to do better. At the local, more proximal level, compatibilism tends to emphasise elements of volition that apparently strengthen our claims of agency, ownership and control. Much of it seems sensible, and a fair deal of it seems to be of a kind with many normative ideals of agency, maturity, and rationality at play in the Western cultural traditions.

But the critical test for a would-be compatibilist, as far as the traditional debate is concerned, involves facing up to the possibility that ours is a deterministic universe; and to then imagine, in various ways and with varying degrees of detail, agents whose lives play out according to the way the deck was stacked before they were born (plus the laws of nature). That is, the real test for compatibilism is to take up a global perspective, and to closely examine lives led in a deterministic universe where every choice and action follows straightforwardly (if in some massively complex and chaotic fashion) from the ‘deck’ of variables as they were set before the lives under examination began, and where every conditional ‘could have done otherwise’ is, from this global perspective, somewhat neither here nor there because there is no causal space for the agent to deviate from the path they in fact followed.

There is much that a compatibilist would like to remind us at this point about what we should and should not infer from such exercises in imagination and perspective-taking. A lot of what

they might say has a great deal going for it – so long as we keep our focus sufficiently local (as opposed to global), and remind ourselves that determinism does not (and ought not to) imply fatalism. And yet in much the same way that Nagel admits or complains about thinking something different about the problem every time he visits it, something worrying lurks in the background and pops in and out of focus as we let the global perspective in a deterministic universe inconsistently influence our intuitions and arguments about freedom in a compatibilist world. On one hand, we are not mere hapless victims – what we do, how we deliberate and choose matters. That is what it is to be a human agent. And yet, on the other hand, we look and feel like the ultimate victims, recognising that a deterministic universe leaves us no room to manoeuvre that has not already been prefigured in the stacking of the deck, and yet persisting (indeed, needing to persist) as if it did.

I am not fond of the liberal use of thought experiments, but one or two might help pin down the kind of sentiment or intuition that the compatibilist needs to be confronted with. For a start, it is helpful to draw an analogy to a rough-and-ready theological version of our current concern: suppose that the universe is deterministic, and that it was created by a suitably empowered supernatural being – let's call her God. In creating this deterministic universe, God lays down the laws of nature and sets the initial conditions at the big moment, time t_1, for everything that follows. In this deterministic universe, for every time t_{1+n} after t_1, there is only one possible state for the universe at t_{1+n} as follows from the combination of the initial conditions at t_1 plus the laws of nature. In the theological version of the puzzle, it sounds decidedly strange for God or her earthly devotees to claim that humans in this universe have 'heaven-or-hell' free will[85], because everything including any token choice or action on the part of a human agent at some time t_{1+n} was, in a significant and freedom-undermining sense, prefigured and *uniquely* constrained by the initial conditions and laws of nature set up by God at t_1.

As far as the implications of imagining a deterministic universe go, God's agency in the above case is as crucial as it is incidental: incidental because there need be no agency involved in setting the initial conditions or laws of nature – the end result for agents is the same; and yet crucial because it brings into focus the strangeness involved in talking about freedom under these circumstances, despite all the good (local level) sense that compatibilists

[85] To borrow an expression from Strawson (1994).

have to offer us about the nature of our agency. If God set the laws and the variables and thereafter things unfold like proverbial clockwork, there just isn't space for free will, whatever things look like at the local, proximal level.

Mele (1995, 2006) considers two arguments or threats to his own compatibilist proposals that, to my mind, represent a secular spin on this much older theological headache. In Chapter 10 of *Autonomous Agents*, he discusses a series of claims that he thinks might be "naively" (Mele, 1995, p189) thought to undermine his proposed compatibilist criteria for psychological autonomy before proposing what he considers a more threatening counterexample. Leaving some finer details aside, Mele's (1995, p187) proposed sufficient conditions for compatibilist psychological autonomy[86] are:

1. The agent is an ideally self-controlled agent.[87]
2. The agent has no compelled*[88] motivational states, nor any coercively produced motivational states.
3. The agent's beliefs are conducive to informed deliberation about all matters that concern him.
4. The agent is a reliable deliberator.

The "naïve" claims against these conditions can be reconstructed[89] as follows (Mele, 1995, pp189):

- Consider a subject *S* whose world is deterministic.
- There is a supremely intelligent being *X* at *S*'s world (who is not *S*) who knows in advance everything that *S* will ever think, decide, intend, and so on.
- This supremely intelligent being *X* is, furthermore, *S*'s creator.
- Under these circumstances, it is claimed that *S* is not autonomous.

Mele (1995) is obviously not persuaded by these "naïve" claims. In his reply to the version of the claim that I have just reconstructed, he counters:

> *X*'s creating an agent whose (entire) future *X* foreknows at the time of creation, including the agent's future decisions, does not suffice for *X*'s being in control of the agent's psychological life. The creator may foreknow that the agent he is about to create will enjoy compatibilist psychological autonomy. (Mele, 1995, pp189-190)

---

[86] We touched on these in Chapter 3 when considering the case of Betty and the basement.

[87] As we saw in Chapter 3, self-controlled individuals "are agents possessed both of significant motivation to conduct themselves as they judge best and of a robust capacity to do what it takes so to conduct themselves in the face of (actual or anticipated) competing motivation." (Mele, 1995, p.5).

[88] For Mele (1995), a 'compelled*' state is one where the compulsion has in no way been arranged by the subject occupying that state. A hypnotic wish not to smoke might have the appearance of a compulsion of sorts, but if the agent arranged for the insertion of that wish into their psychology by way of voluntary hypnosis, it is not an ordinary compulsion – hence, the agent is not in a compelled* state.

[89] My reconstruction glosses over the initial two claims he discusses, and instead combines the key ingredients of each claim that contribute to the 'strongest' third claim.

Before we turn to what Mele does regard as a more serious threat to his account, a number of comments are warranted at this point. First, there is a distinct echo in Mele's last sentence of the would-be theological defence of free will we've just encountered – substitute 'free will' for 'compatibilist psychological autonomy', and Mele doesn't sound like he is saying much different in spirit from what God and her earthly servants had to offer in the theological case. This is, in one sense, unsurprising, given the complicity of all-knowing creators in both cases. But what is surprising is Mele's idea that he could have offered an account of compatibilist psychological autonomy that can so easily escape what God and her minions could not.

There are two key ideas in Mele's dismissal of these "naïve" claims. First, he reminds his opponent that he is offering only conditions for compatibilist psychological autonomy, and thus that these conditions cannot be brought into doubt because *S* does not satisfy some set of conditions for incompatibilist autonomy. Second, he seems to think that a lot hangs on the claim that *X* is not *in control* of *S*'s psychological life just because *X* created *S* and knows everything that *S* will do in advance of its happening. That is, Mele attributes great significance to the absence of *proximal* control (and intervention) by *X* in *S*'s life. I will comment on each idea in turn.

Mele's insists that because his proposed conditions were only offered as sufficient conditions for compatibilist psychological autonomy, they cannot be undermined or criticised because an agent that satisfies these conditions does not also satisfy incompatibilist conditions for such autonomy. In terms of Mele's larger project in *Autonomous Agents*, that might seem fair enough to some. But Mele's response here helps highlight the kind of frustration and lack of possible movement that so often characterises the traditional debate. What the imagined "naïve" respondent is complaining about seems simple enough: if Mele thinks he has provided sufficient conditions for psychological autonomy, he has not, because in a deterministic world *S* would, according to the respondent, lack autonomy. In other words, what Mele has offered as (compatibilist) sufficient conditions for psychological autonomy are not convincing enough for the respondent to agree that *S* possesses something worthy of calling psychological autonomy in the deterministic world that *S* occupies. The bulked out version of the complaint (including our *uber*-agent *X*) can be seen as a way of bringing the source/s of dissatisfaction into greater focus. And yet Mele thinks he can dismiss the basic complaint as being based on a naïve misunderstanding of some kind – that the respondent is

clearly operating with an incompatibilist notion of psychological autonomy, so their objection can't touch Mele's compatibilist conditions.

Part of the problem here is fundamental to the way in which the field of play is defined within the traditional debate. From a compatibilist point of view, any conception of freedom (or free will, free action, free agency, psychological autonomy, autonomous action, or autonomous agency) that still imagines some kind of threat being posed by *S*'s world being deterministic is likely to be dismissed out of hand as an incompatibilist notion of freedom (*et cetera*), and as thus irrelevant to the evaluation of any token compatibilist account or conception of freedom. Of course, a compatibilist can legitimately expect objections based on these alternative conceptions to be held off for long enough so that a compatibilist account can be put forward and defended without itself (*qua* compatibilist) being dismissed out of hand. But when that opportunity has passed and the compatibilist account must face its critics, it is unhelpful, unenlightening and cognitively stultifying to the then dismiss those critics *qua* incompatibilists. After all, as far as the traditional debate goes, it is committed libertarians and hard-determinists who deserve to be labelled 'incompatibilist'; 'neutral' parties to the debate who remain unconvinced of the virtues of a token compatibilist account are not 'actively' incompatibilist in any meaningful sense.

What I think this points to is a tendency within the traditional debate (but especially within compatibilist circles) to be excessively dismissive of lingering opposition and criticism – even when coming from nominally neutral camps – based on the following kind of logic: (a) compatibilist dismissal – "My account has clearly demonstrated how the only freedom (autonomy, etc.) worth wanting is entirely compatible with the truth of determinism; if you can't yet see that, or you are not satisfied, that represents some kind of cognitive failure and incompatibilist cognitive intransigence on your part."; (b) libertarian dismissal – "It is obvious that we are free, and my account has made it clear the form that our libertarian freedom must take in order to do justice to our freedom; if you can't yet see that, and you are willing to settle for some poor compatibilist substitute, or live in denial of your freedom as a hard determinist, then you are suffering from some kind of cognitive failure."; or (c) hard compatibilist dismissal – "My account has made it obvious that compatibilism is wrong, and that libertarianism is misguided and doomed to failure – we are not free; if you can't see that, then you are living in compatibilist or libertarian self-delusion, and you are suffering from some kind of cognitive failure."

Once again, such dismissals need not be in any way unique to this particular philosophical debate. Philosophers love to dismiss their critics as suffering from some combination of (i) a willing inability to properly understand plus (ii) a blind dedication to some opposing position. The claim is not that this is unique, but only that such dismissal is particularly rife in this well-trodden, extensively mapped-out, high-stakes debate over free will.

Perhaps Mele does not think he is guilty of any such dismissal, however. Perhaps his mock-speculation that all-knowing *X might* foreknow that he is about to create a being possessed of compatibilist psychological autonomy is supposed to be well-grounded by the second key idea I mentioned – namely, the idea that the absence of any *proximal* control by *X* over *S*'s psychological life is crucial to blocking the inference from *X*'s creation of *S* and his foreknowledge of *S*'s life course, to *X* somehow having control over *S* in an autonomy-threatening sense. We know how this kind of argument is supposed to work because it is more or less standard compatibilism: look at the local and proximal level of the agent acting in the world, and if there is no compulsion and no interference, then there is no problem. The existence and epistemic prowess of *X* doesn't change that. Or so the story is supposed to go.

But for the neutrals and undecided participants in the debate – not to mention the committed incompatibilists – *X*'s existence and nature do make a difference. It is a bit like imagining a deterministic flip of a coin played out in extreme slow motion: you can put as much apparent time and distance between the person who flicked the coin into the air and its eventual position heads or tails up on the ground, and yet the flicker of the coin still plays a defining role in determining the result of the toss. Similarly, when you imagine (or, I suppose, believe in) the existence of a being like *X* who sets the initial conditions and the rules by which events will unfold, the worry remains that an absence of obvious proximal interference and/or control does not alter the relevant distal facts of how any particular choice or action of *S*'s came about. That is, we see the global perspective once again coming back to haunt the compatibilist, despite their best efforts at the local, proximal level.

That there is a threat contained within this "naïve" objection is, I think, borne out by the way in which Mele characterises what he does regard as "an apparently serious threat" (Mele, 1995, p190). Mele asks us to consider a modified version of the case we have been considering. In the modified case – let's call it X^2 for short – we are asked to imagine the

creator *X* creating an adult agent whom Mele calls Fred. *X* creates Fred precisely because *X* wants a certain event, *E*, to occur a year later. Mele further specifies that we are to imagine (using the conceptual machinery he has already introduced) *X* endowing Fred with a set of *sheddable* desires and values[90], in the knowledge that giving Fred these desires and values now will partly determine that Fred will, a year down the line, deliberate on and decide to do *A*, where his *A*-ing will bring it about that *E* occurs. It is *X*'s understanding of the complex deterministic web in which Fred's life over the next year will play itself out that leads *X* to give Fred precisely the batch of (sheddable) motivational states that he endows him with. If, as Mele asks us to imagine, we allow that Fred could satisfy all of Mele's proposed conditions for compatibilist psychological autonomy, we might then be tempted to infer that Fred nevertheless lacks compatibilist psychological autonomy despite satisfying these conditions.

Mele thinks that the intuition guiding us towards this conclusion is one that frames Fred as, in some significant sense, an *instrument* of his creator *X*; and so because he is an instrument whose internal psychology has been tuned in such a way as to produce, at a distance, a particular result (his *A*-ing which brings about *E*), he lacks psychological autonomy. Mele thinks this is, at least, a compelling view of the case for libertarians. But Mele does not think this intuition will be shared by compatibilists. He reminds us that Fred has no compelled motivational states in this scenario; nor does *X* compel Fred to make any specific deliberative judgements or choices, nor to perform any actions. Fred can do all the things that compatibilists want to emphasise in the case of agents who have not been created in the way that Fred has: in a compatibilist sense of 'can', Fred can reflect on his motivations and values, he can change or shed those values, and he can choose or form new ones. Of course, he won't do any of these things – he will *A*, thus bringing about that *E* – and it is causally determined that he won't do any of them, and that he will *A*. But that, for Mele, is just what you get when you assume that compatibilism is true.

In addition to this bit of compatibilist bullet-biting, Mele further recommends revisiting the idea of the Fred's creator *X* being a *compatibilist creator* who somehow values the fact that

[90] I will not go into great detail on Mele's notion of sheddable and unsheddable desires and values. Unsheddable (or, more carefully, *practically unsheddable*) values are ones that, over a given time period, an agent is in effect stuck with.because of the way in which these are entrenched within the agent's psychology. While the value could be removed or replaced in principle, Mele's idea is of a value at a given point (or over a given period) in an agent's life whose removal "is not a psychologically genuine option" (Mele, 2006, p167). See Mele (1995, pp149-173; 2006, pp166-170).

Fred will bring about *E* in an autonomous fashion. It is not clear exactly what point Mele wants to make in this regard, but he seems to allow that, somehow, a shift in perspective to that of a compatibilist creator who values autonomous agents acting autonomously might encourage suitable rival intuitions to those he attributes to the committed libertarian.

The experiment in perspective-taking turns out to be a failure. Contemplating the point of view of a 'compatibilist creator' to whom it matters that the things they want to happen (*E*) are brought about by 'autonomous' agents acting 'autonomously' only reinforces an unfortunate link back to theological versions of the free will problem. If God sets the variables and makes the rules, and things are then all causally determined to happen in one and only one way, neither God nor our claims to having free will look very good when considered in the context of some 'heaven-or-hell' final judgement. If, for some mysterious reason, God is a compatibilist and it matters to her that her pawns play out her cosmic game according to the rules that define 'compatibilist autonomy', it doesn't make it any less of a cosmic game, our freedom of will any less of a sham, or the final judgement any less of a gross and catastrophic miscarriage of justice. It is hard to see, therefore, how compatibilism can benefit from any association with this theological version of the debate.

What is equally frustrating in Mele's discussion (and it is, again, symptomatic of many parts of the traditional debate) is that he sets out his comments on this case as if it could not possibly be decisive in swaying one between compatibilist or non-compatibilist points of view. What will make the image of Fred as an instrument compelling, it is suggested, is a pre-existing commitment to libertarianism, rather than any merits of the example and truths it might claim to point to. This treatment of the example is, in an important sense, resigned to a stalemate and/or immovable positions, instead of pressing the case to test out its full merit and import.

This task has, however, been taken up in the interval between *Autonomous Agents* (1995) and Mele's (2006) *Free Will and Luck*. Here, Mele raises the stakes by presenting what he calls *the zygote argument* – what he describes as a manipulation argument for incompatibilism. Following Mele, we can develop the argument in two steps – first with some scene setting, and then the argument itself. In the scene setting, we find a creator Diana at work in place of the nameless *X*:

> Diana creates a zygote *Z* in Mary. She combines *Z*'s atoms as she does because she wants a certain event *E* to occur thirty years later. From her knowledge of the state of the universe just prior to her creating *Z* and the laws of nature in her deterministic universe, she deduces that a zygote with precisely *Z*'s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to *A* and will *A* on the basis of that judgment, thereby bringing about *E*. If this agent, Ernie, has any unsheddable values at the time, they play no role in motivating his *A*-ing. Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly exercises his powers of self-control and has no relevant compelled or coercively produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator. So he satisfies a version of [Mele's] proposed compatibilist sufficient conditions for having freely *A*-ed. (Mele, 2006, pp188)

Needless to say, in Diana's deterministic universe, Ernie does *A* thirty years down the line, thereby bringing it about that *E* occurs.

With the scenario in place, Mele (2006, p189) formally states the premises of the zygote argument as follows:

1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
2. Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
3. So determinism precludes free action and moral responsibility. (Mele, 2006, pp189)

There are a number of important differences between this argument about Ernie and the earlier case involving Fred. The primary difference is that, unlike Fred, Ernie is not miraculously created *de novo* as an adult agent with various psychological states that will be critical to his *A*-ing. In fact, what we are asked to imagine is a scenario that does not involve Diana in any kind of direct creation of, or tampering with, Ernie's psychological make-up. Diana only deals in the atomic composition of the zygote *Z* which, having been slotted into the deterministic matrix of the universe, will develop into Ernie who, thirty years later, will *A* such that *E* comes about.

As Mele notes, this change in the creation scenario helps block any externalist objections which, in a case like Fred's, might hold that his creation *de novo* is a conceptual impossibility because his psychological states lack the relevant kind of historical links required to fix the content required for the imagined scenario to play out. Ernie's development, in contrast, offers all the normal opportunities for content-fixing interactions with the environment (and others) that are available to other agents whose zygotes came into being in biologically normal ways. Indeed, nothing about Ernie's development, bodily or psychological, differs from normal cases following Diana's initial intervention.

Although Mele does not explicitly make this point, the apparent absence of anything that looks like specifically *psychological* manipulation also helps block any attempt to dismiss the case as some kind of analogue of brain-washing. Of course, Diana knows how Ernie's psychology will develop – it is her ability to foresee the relevant complex patterns that allows her to deduce the appropriate atomic arrangement of *Z*, given the prior state of the universe and the laws of nature, to ensure a psychology that will bring it about that he *A*'s in thirty years' time. But no specifically psychological states are created or manipulated in Ernie.

It is, I think, both this lack of specifically psychological interference, along with the distance in space and time between Diana's manipulations and Ernie's *A*-ing, that makes Mele's zygote argument particularly effective; and this is especially so with regards to premise 2. The nature of Diana's manipulation is such as to leave compatibilist and incompatibilist alike with no obvious candidate differences between *Z* and any normal human zygote that could be relevant to issues of free action and moral responsibility. Moreover, as Mele (2006, p190) himself highlights, premise 2 manages to capture some of the spirit of the famous consequence argument, without requiring any prior commitment to incompatibilism. That is, Mele thinks that while compatibilists cannot accept the consequence argument (which is, after all, an argument for incompatibilism), premise 2 is consistent with compatibilism, because the idea that there is no salient difference in freedom or in moral responsibility between *Z* and a normal human zygote is consistent with Ernie having acted freely and with moral responsibility for his *A*-ing.

So the likely impact of the zygote argument turns on responses to premise 1. Mele (2006) thinks that judgements about premise 1 will vary, first, based on pre-existing commitments to compatibilism or incompatibilism. This much, then, is in common between the cases of Fred and Ernie – it is held that these cases won't shift the ground beneath those who have independent commitments to one or other side of the traditional debate. But unlike his discussion of Fred's case, Mele (2006) now seems to allow that there is room for what he calls (p190) "*reflective agnostics*" to play some role in adjudicating on the zygote argument. And his speculation is that such agnostics' take on premise 1 will be strongly related to the extent to which they have already judged the prospects for indeterministic varieties of agency to deliver up free action and moral responsibility.

As for Mele himself, given his avowed agnosticism about compatibilism together with his double-layered account[91], he recognises in the zygote argument an important test case for him to address explicitly as regards his commitment to compatibilism. As he puts it, "even patient readers may want me to put my cards on the table and say whether *I* believe that a full-blown version of the zygote argument would show compatibilism to be false" (Mele, 2006, p191, italics in original). And after a few more pages of scene setting, he finally puts his cards on the table:

> Premise 1 has some intuitive pull on me, but not enough to move me to accept it. I am agnostic about premise 1, as I am about compatibilism. (Mele, 2006, pp194)

So much, then, for putting one's cards on the table. In the end, Mele has combined his die-hard agnosticism about compatibilism with a somewhat traditional tendency towards inertia, and has let the zygote argument have its airtime without having any decisive influence on his position.

Without making any radical claims of novelty on behalf of the zygote argument, it does strike me as being peculiarly effective in discouraging fence-sitting amongst both the committed faithful and the 'reflective agnostics'. It is, in many ways, a secularised version of the theological scenario discussed earlier – an all-knowing, all-powerful creator of a deterministic universe, amongst whose creations are creatures who think they are free, and free because the creator made them so (and because it matters to the creator that they are so, etc.). Those bothered by this theological puzzle are presumably, amongst other things, unsatisfied with the idea that any of this could or should matter to the creator. The creator set the conditions and the laws, and nothing could have in actual fact turned out differently given those initial decisions and interventions, so how could it possibly matter one way or another (except on the basis of some mere whim or trivial preference on the part of the creator) as to the pattern they want played out in this particular cosmic game?

What is absent in the theological case, but brought into the zygote argument, is deliberate and specific intent on the part of the intervener to make a certain thing happen – that is, the zygote argument (but not the theological puzzle) explicitly builds in an element of manipulation. Whether this makes too much difference in the final analysis is open to debate – all-knowing creators could perhaps, with a few inferential steps, be accused of choosing/

[91] As will be recalled from Chapter 3, Mele (1995, 2006) thinks that he can offer a series of compatibilist conditions for autonomy (1995) or free agency (2006) which can then be supplemented with an additional incompatibilist condition as necessary.

willing each and every event that unfolds in the universe for which they set the initial conditions and the laws. Be that as it may – the zygote argument offers a secularised challenge to compatibilist orthodoxy by presenting an apparently unambiguous case of active[92] agent manipulation in which we are asked to share the following intuitions: Ernie *A*-ed because Diana wanted his *A*-ing to bring it about, at a particular point in time, that *E* took place; whatever else may be true of Ernie, it is true that (in the deterministic universe he inhabits) he was causally determined to *A* and nothing else, thus fulfilling Diana's plan for her pawn, and precisely because she had this plan for him; so, despite any (compatibilist or other) appearances to the contrary, Ernie's *A*-ing was not something done of his own free will. But of course, with these intuitions in place, we are invited to recognise that there is nothing unusual about Ernie's *A*-ing other than a distant historical fact about the atoms in his zygote, and Diana's plans for him. So, if Ernie did not *A* freely, and we can't find any other distinguishing feature of his *A*-ing that might disqualify it from being free, then all our *A*-ings are also not free if we inhabit a deterministic universe.

Perhaps incompatibilists would claim that the causal straightjacket that is the deterministic matrix into which Diana slots Ernie's zygote is just what they have been trying to highlight all along – perhaps even more directly – with arguments such as the consequence argument. But as we have seen, Mele and others are inclined to think that too many of these arguments and problem cases require a prior commitment to incompatibilism before they can get any purchase in swaying opinion. The apparent virtue of Mele's formulation of the zygote argument is that a significant part of it – premise 2 – requires no such prior commitments; this leaves us to focus on premise 1 and reasons that one could offer for contesting it.

Mele, as we have seen, wants to remain agnostic about it, but is this sensible? Agnosticism in this instance suggests trying to hedge one's bets on compatibilism a little too far (just in case our universe turns out to be deterministic, or just in case libertarians fail to make a sensible case for indeterministic agency). Mele (2006, p193) acknowledges that cases like that of Ernie "prevent [him] from flatly endorsing compatibilism", but this is surely to do the argument an injustice. The zygote argument should either force a bit of committed compatibilist bullet-biting (for which I suspect Dennett would volunteer), or it should force one to confront a deep incompatibilist intuition in oneself – namely that, in the end, free

[92] That is, 'active' in a way that contrasts with cases of non-intervening controllers who could actively intervene, but (contingently) do not.

agency just is not possible in a deterministic universe because Ernie *qua* agent has everything a normal agent in such a universe has in order to qualify as free, and yet in the end he is not because he is Diana's plaything or tool.

What should we make of Mele's speculation that intuitions or opinions about premise 1 are most likely contingent on an individual's views concerning the compatibility of indeterministic agency with free and morally responsible agency? It is, of course, quite possible that he is right about this, and that responses to the zygote argument will not be independent of considerations of the available alternatives, especially libertarian ones. Someone who, like Dennett, has persuaded themselves of the utter misguidedness of all conceivable indeterministic alternatives to compatibilism and yet (like Mele) remains convinced that our claims to free agency are more justified than any particular account that has been given of free agency, will no doubt feel that they have to deny premise 1, and convince themselves that some erroneous inference just must be hidden in our thinking about this premise, even if we can't say where.

And yet, for all this, premise 1 seems remarkably clear. Ernie is not a free agent and he is not morally responsible because his *A*-ing was something deliberately set in motion by the prenatal interventions of Diana. There are no non-intervening controllers here – this is straightforward manipulation (although at a distance). And yet the nature and timing of the intervention is such as to be largely indistinguishable (in terms that might matter to Ernie's agency) from the arrangement of atoms in normal zygotes by 'blind' forces of nature. The intuition that should be encouraged here, in the uncommitted and the agnostic, is that *if this is what compatibilism allows, then compatibilism has settled for something short of free will*. Free agents cannot be the play things of gods and other intervening manipulators; and to the extent that compatibilism cannot explain away a case such as Ernie's, compatibilism will have sold us short in the free agency stakes. No agnosticism, then, only bullet-biting or a basically incompatibilist conclusion that we cannot, after all, be free agents in a deterministic universe.

Where, then, are opinions about the compatibility of indeterministic agency and free agency supposed to have their influence on our responses to the zygote argument in general, or premise 1 in particular? In my own case, I have made it clear that extant libertarian accounts of indeterministic agency are not promising at all, at least in so far as they insist on inserting

indeterminism into moments of volition. Yet I think premise 1 of the zygote argument is true, independently of such considerations. Freedom that can be manipulated in this way by god and other intervening manipulators just isn't what I had in mind when thinking we might have free will. It seems better to face up to something like Nagel's state of deep puzzlement than to simply insist that premise 1 *must* be flawed just because libertarianism seems so unpromising.

More generally, we need to be wary of certain cognitive tendencies that risk turning one form of (arguably sensible) philosophical caution into an unjustified philosophical orthodoxy of optimism. We should be suitably cautious in pronouncing phenomena and our beliefs about them as illusory just because we are able to frame some vexing philosophical puzzle (especially of a sceptical variety). Yet we should also recognise a tendency towards conservative positions and orthodoxies that are significantly motivated by a combination of philosophical optimism and wished-for empirical insularity. Orthodoxy in the philosophy of mind tends towards varieties of non-reductive materialism in part because these optimistically promise to allow us to save most of what we want to say about the mind while at the same time insulating such claims from threat by present and future empirical findings. Orthodoxy in the traditional free will debate tends towards varieties of compatibilism in part because these optimistically promise to save one or more varieties of freedom 'worth wanting' while at the same time insulating this freedom from a wide range of present and future empirical threats, including possible discoveries about the extent to which our (macroscopic) universe operates deterministically. By contrast, eliminativism and reductionism in the philosophy of mind, and libertarianism in the traditional free will debate, can be seen as more conceptually and intuitively disruptive; and they very often require (or are subject to) elements of empirical support or refutation that make them cognitively more open-ended than the anything marked by the empirical insularity just described.

My suspicion is that the non-committed or agnostic who reject premise 1 out of concerns related to the likely compatibility of indeterministic agency and free agency might be manifesting this kind of conservative optimism in the face of what is otherwise an intuitively unambiguous case. It is not just, as Mele (2006, p191) surmises, that this group "either believe that there are free and morally responsible agents or are more inclined to believe that than they are to believe that there are no such agents", but that they further want to insulate such beliefs from all conceivable conceptual (problems with indeterministic agency) and

empirical (the universe and/or the brain might work deterministically) threats. Compatibilism has always made these kinds of optimistic promises, and offered refuge. Unlike Mele[93], I think the zygote argument provides a relatively fresh (if not necessarily novel) means to exposing a critical lingering deficiency in the defences of this presumed refuge.

If I am right in thinking that the uncommitted (or the reflective agnostics) should recognise the flaw in compatibilism highlighted by the zygote argument, what of compatibilists?[94] As we have seen, Mele has been careful in arranging the case such that there is no obvious route by which a compatibilist can exclude Ernie from a claim to being free and responsible in his *A*-ing within a compatibilist framework. He suffers from nothing resembling compulsion or coercion; he has not been brainwashed (*Z* has no brain yet) or psychologically manipulated (*Z* has no psychology yet); he is not subjected to any kind of abnormal proximal (or, for that matter, post-natal) interference; unlike Frankfurt-type examples, Ernie's case does involve active intervention rather than the mere possibility of intervention that, contingently, is never actualised; and Diana's interventions are (in so far as it matters) targeted at sheddable motivations and values such that Ernie qualifies as having (conditional) alternative possibilities for choice and action just like any other compatibilist agent.

One potential line of response would be to challenge the plausibility and soundness of the thought experiment on which the zygote argument is based – perhaps along lines similar to those pursued by Dennett (1991) in his discussion of the implausibility of the brain-in-a-vat thought experiment. Dennett (sensibly) highlights the astronomical complexity involved in trying to mimic *all* of the signals being received by the brain over even a small time slice, once one remembers to include all of the *internal* inputs from the body, the interactions between these and changes in signals from the world, plus trying to mimic all the neural signals that would be the registerings of changes in the brain's own internal milieu (esp. its blood flow and chemistry). The requisite computational complexity to sustain any kind of world-embedding illusion could not conceivably be handled by the most powerful computers

[93] Mele (2006, p191) thinks that "the addition of Ernie's story to the collection of relevant things reflective agnostic have reflected on [in thinking long and hard about free action and moral responsibility] would not move many of these people out of the conjectured majority [who doubt the truth of premise 1] and would not – at least very quickly – give many of them a significantly brighter view of indeterministic agents' prospects for free action and moral responsibility."

[94] That is, obviously, compatibilists who are not opting for the bullet-biting response that flatly rejects premise 1.

we have imagined.[95] Dennett suggests – plausibly, in my view – that imagining a brain-in-a-vat at this level of detail should persuade us of the extreme implausibility, and thus dangers, of grounding or testing important ideas about the mind and brain using this particular intuition pump.

Similarly, I imagine that a compatibilist might try to object to the zygote argument's underlying thought experiment on the grounds that, when imagined in sufficient detail, its demands would stretch the bounds of possibility beyond credulity. On the surface, Diana's activities seem 'simple' enough: they involve the arrangement of atoms in a microscopic single cell that has formed from two human gametes. As far as the thought experiment goes, no other kind of intervention in any other part of the universe, or at any other time, is required. Thus far, at least, the thought experiment seems safe from the kind of combinatorial explosion of computations and interventions/ inputs that unfolds *over time* in the brain-in-a-vat example. But this appearance is deceiving.

The combinatorial explosion of computations in the brain-in-a-vat case is dynamic and unfolds over time in a way that, in principle, allows for ongoing adjustment and error correction of both interventions and future predictions.[96] Diana's challenge is, in contrast, primarily one of performing an astronomically complex prediction that then needs to be distilled down into a series of hyper-fine-tuned values for a one-off, make-or-break arrangement of atoms in the zygote *Z*. This raises three possible levels at which one could challenge the conceivability and plausibility of this scenario.

First, the size and scope of the predictive task facing Diana is impossibly massive, involving not only gargantuan chunks of the deterministic causal matrix into which *Z* and (later) Ernie must be slotted, but also the dynamic interplay between *Z*/ Ernie and all of the environments it/he will occupy over the next 30 years. Second, we have no conception of what would be

[95] Dennett (1991) describes the problem as one involving a combinatorial explosion. See, especially, pages 4-7 of *Consciousness Explained*.

[96] A post-Dennett embellishment of the scenario might allow, for example, that the computer system can utilize some degree of *ex post facto* error correction by manipulating memory traces in such a way that the brain-in-a-vat is not bothered by, and does not in any way dwell on, small discrepancies, inconsistencies, or other shortcomings in the virtual world with which they are presented. Another embellishment might allow or highlight the possibility for running multiple ongoing parallel computational simulations of the brain-in-a-vat under varying inputs – there being no limit on the computational resources available to the machinery that can be put to use – such that, at any instant, the controlling computer already has available the results of various simulations to utilize and potentially switch between when it comes to choosing the next period of input and dynamic interplay between the brain and its virtual environment.

involved in distilling out the relevant facts about atomic arrangements within *Z* that will provide suitable degrees of tolerance on Ernie developing the relevant psychology that, under the predicted circumstances in 30 years' time, will result in his *A*-ing. Moreover, there are many considerations from within the philosophy of mind that would suggest the impossibility of such a series of backwards inferences. And third, even allowing that the aforementioned could be successfully tackled in principle, there would remain questions about the required degrees of accuracy in both predictions and manipulations that could make what would be, at best, a highly chaotic if deterministic system spit out the right consequences 30 years down the line, given what we know about chaotic systems and the differences that can be made by the smallest of differences in initial conditions within such systems, plus the absence of any opportunity for further intervention and error correction. As is arguably always the case, postulating a supernatural being to do all the important work really is, on reflection, a thoroughly misleading (if historically popular) way of imagining things.

As someone who is generally sceptical about the value of many philosophical thought experiments, I am very sympathetic to the above critique of the Diana-Ernie scenario that underpins the zygote argument. But my views on thought experiments in general are not at issue – it is the plausibility of the above objections within the context of the traditional debate, and more specifically as a means for a compatibilist to side-step the zygote argument, that requires consideration.

Many compatibilists cannot legitimately lay claim to this sort of 'inconceivability' response because they are already committed to the value and utility of equally preposterous thought experiments, especially those directed at exposing the alleged flaws in libertarianism. Molecule-for-molecule doppelgangers are, at best, slightly more plausible philosophical test cases than Diana's amazing feats of science and engineering, in part because they aren't supernatural *ex hypothesi*[97], and because we don't usually pause to ask anything about the origins of these creatures[98]. Asking such questions (while refusing to allow non-answers,

[97] Kind of. Barring the inventiveness of Star Trek's creators, would we not think the idea of a body scanner-generator to be something supernatural? I suspect we could only see this idea as 'natural' based on a series of imaginative failures.

[98] Given the differences, noted by Edelman (2004, Edelman & Tononi, 2000) and others, that we can expect at the level of anatomical structure within the brains of genetically identical twins, there is no plausible alternative to Star Trek-like scanner-generator machines when it comes to the business of producing doppelgangers, and most of us only succeed in imagining those (to the extent that we really imagine them at all) because of a TV programme.

such as positing the existence of some of the molecule-for-molecule scanner-generators that populate certain debates about personal identity) will quickly lower doppelgangers into the same murky world of impossible and potentially misleading scenarios that do cognitive work for us for just as long as we fail (actively or passively) to imagine them at relevant levels of detail.[99] But doppelgangers who exit an indeterministic choice situation with different outcomes are, for the compatibilist, such an important tool in exposing the alleged flaws of libertarianism that they can ill-afford to disqualify 'in principle' thought experiments of either the doppelganger or Diana-Ernie varieties on grounds of their preposterousness in practice.

Presuming, then, that Mele is right about premise 2 being consistent with and acceptable to compatibilists, it is hard to see how a compatibilist can do anything other than bite the bullet, deny premise 1, and insist that despite any 'misleading' intuitions to the contrary, Ernie *is* a free agent who *is* morally responsible for his actions, including his *A*-ing. And so much the worse for compatibilism.

Perhaps this accusation of bullet-biting is too quick. Or, rather, perhaps a committed compatibilist would want to contest the sense in which they are supposedly biting the bullet, as opposed to remaining firm and comfortable in their compatibilist convictions because the zygote argument has not offered them any persuasive reason to question these convictions. How might such a compatibilist respond to the argument?

An initial response might go something like this[100]. *Ex hypothesi*, the manipulator (Diana) in the zygote argument scenario must arrange things such that all relevant the conditions for compatibilist autonomy (CA) are respected. (Depending on which or which kind of compatibilist is responding, exactly what these conditions are may vary depending on their preferred account of CA. For present purposes, however, it should do to just keep Mele's conditions for CA in mind.) For example, in Mele's version of the argument, Ernie must come to act as he does under the influence of sheddable (as opposed to unsheddable) values, he should deliberate normally, etc.; and he should not be under the sway of any 'innate' programmes or magical psychological 'time-capsule' of some kind that suddenly kicks in after 30 years to produce his *A*-ing.

[99] For more on the risks and problems of many philosophical thought experiments, see Dennett (1996).
[100] I am grateful to Mark Leon for the outline of this and the following compatibilist response to the zygote argument.

But then it seems open to a compatibilist to respond that, given an independent case that has been made for their preferred account of CA, they are untroubled by the case of Diana and Ernie because if the conditions for CA have been respected, then nothing has transpired in this case to create a concern over Ernie's freedom or responsibility as an agent. Premise 1 is, thus, false. And this should not be mischaracterised as some form of intransigent bullet-biting. Instead, what we have here is exactly what we should expect from a compatibilist who truly understands and values the conditions laid out in their preferred account of CA: that is, when these conditions have been satisfied, we have all that we should reasonably want in order to secure claims of freedom and responsibility.

There is a straightforward counterargument to this initial compatibilist response. Compatibilism is, after all, supposed to (historically) champion notions of negative freedom, most obviously including freedom from certain kinds of interference, including deliberate manipulation. Surely it is incontestable that Diana's manipulation of Ernie's zygote *Z* in order to serve her own ends is an unambiguous case of (remote) interference, even if the apparent absence of any (proximal) psychological or other interference might make Ernie's *A*-ing closely resemble his freely *A*-ing under suitably different circumstances? The idea that conditions for CA do not successfully pick out those relevantly different circumstances is supposed to be part of the point of the argument. We have a clear case of manipulation (what else could it be?), which is precisely the kind of thing that compatibilism should find problematic. The challenge is then to say how, if at all, such manipulation could *not* be threatening to autonomy and freedom, given that conditions for CA have already been satisfied. To simply reassert that the case does not involve a threat to freedom because conditions for CA have been satisfied is question-begging.

Unfortunately, accusations of begging the questions are just as easily made as they are reversed: our compatibilist respondent might just as well claim that it is me who is begging the question against their preferred account of CA because, faced with a case like Ernie's, I am not satisfied that his claims to freedom and responsibility survive *precisely because* his CA is unscathed. Compatibilism is, historically, not only the champion of negative freedom – it is also home to those who would offer us sufficient conditions for freedom and responsibility *despite* any (misleading) appearances suggesting that freedom is incompatible with, say, determinism, or the absence of alternative possibilities, etc. Perhaps Mele has

shown that we should add CA-respecting zygotic manipulations to the list of potentially misleading scenarios under which we might come to doubt an agent's freedom. To insist that he has done more than this is to beg the question against CA.

Accusations of question-begging are, in this case, unlikely to help matters. Let us, then, consider a more constructive and detailed compatibilist response. Suppose we are asked to consider an analogy between the Diana/Ernie case and a different kind of case in which a CA agent comes to do the will of another. Imagine, for a moment, that Diana's laboratory interventions with *Z* were intended to ensure that (a CA) Ernie would save a drowning child in 30 years' time. Now we are asked to compare this case to that of Sarah, who is an ordinary agent with no superhuman predictive powers nor any special biochemical knowledge, who is aware of the possibility that a drowning child could be saved by the appropriate actions of another (ordinary) agent, Percy. Sarah urgently engages Percy in a bout of CA-respecting rational persuasion, the end-result of which is that Percy decides to act to save the child, and does so.

Our compatibilist respondent would like to propose that, if we are suitably mindful of Diana's commitment (*ex hypothesi*) to respect Ernie's CA in bringing it about that he saves a child 30 years down the line from Diana's interventions, we should come to recognise that there is no *relevant* difference between the Diana/Ernie case and the Sarah/Percy case that would render one a freedom-threatening *manipulation* into saving while the other involves freedom-respecting case of rational persuasion to save a child.

Of course, there are many differences between the two cases, including the obvious fact that Diana's interventions are of a completely different order of complexity to Sarah's; and the fact that Diana's own (i.e. direct) interventions do not include any process of rational engagement with or persuasion of the intended target of her intervention. Nevertheless, the compatibilist's claim is that, in so far as both interventions are part of bringing about the saving of the child by way of an understanding of how each agent (Ernie and Percy) could come to act in the relevant way without in any way threatening or undermining their CA, the intervention of Diana *should not be seen as any more or any less manipulative* (in a freedom- or responsibility-threatening sense) than the intervention of Sarah. And if it is not a case of freedom-threatening manipulation, then there is no onus on the compatibilist to find a means

to distinguish Ernie's case from Bernie's case (where blind forces 'arrange' the atoms in his zygote); and the claim that CA is sufficient for freedom can stand.

A first response to this analogy-based argument is to note that the cases, as presented, have been coloured by the introduction of an ethically-charged scenario: Ernie's generic *A*-ing has become an action or series of acts in pursuit of the goal of saving a drowning child, and Diana's desired event *E* is now (presumably) to be identified with the saving of the child. For the case of Sarah and Percy, this seems a sensible and innocuous enough choice. One may as well improve the likelihood of success of Sarah's attempts at rational persuasion (and reduce the risk of mistaking this for some kind of manipulation of Percy) by making her object an ethically-desirable one. What is less clear is whether or not the ethically-charged example is equally innocuous as content inserted into the vehicle of the zygote argument. Subscribers to an asymmetric account along the lines of Susan Wolf's (1990) might think it a perfectly sensible choice, since Ernie is supposed to be imagined as deciding it is best for him to *A*. But I take it that a generic compatibilist response would not wish to be dependent on Wolf-type asymmetries when it comes to the possible range of free actions.

The question that needs to be pressed is this: does the introduction of ethically-charged content into the zygote scenario make a difference over, say, imagined scenarios involving ethically neutral, trivial, or even ethically problematic content? In one sense, at least, it should not matter, just so long as it is remembered that we could substitute in any of the frivolous, trivial, playful, naughty, or unethical things that an ordinary agent like Ernie might reasonably decide to do of his own free will. Where it does matter, however, is to the success of the analogy that our compatibilist would like to draw between Ernie and Percy.

Let us recall the details of the second case. Sarah, we were asked to imagine, must engage Percy by way of CA-respecting rational persuasion in order to get him to save the drowning child. If we generalise this case, then Sarah needs to rationally persuade Percy to *B* such than an event *F* will come to pass, while respecting Percy's CA. Exactly what *B* and *F* could be will depend significantly on Percy's character and circumstances, because these will influence the likelihood of Sarah's success in persuading him to *B* such that *F* without in any way compromising his CA. Thus, depending on the details, the requirement that Percy be rationally persuaded by Sarah in a CA-respecting manner restricts the range of possible

candidates for *B* and *F*, as well as the relationship between *B* and *F*[101], relative to Percy's character and circumstances. (Which means that, so long as Percy is not some moral reprobate, persuading him to save a drowning child seems a good choice as a case.)

Do any such restrictions apply in the case of Diana and Ernie? It is not immediately obvious that they do apply. Diana's intervention is only at the level of arranging atoms in Ernie's zygote *Z*. She certainly does not engage in an attempt at rational persuasion with adult Ernie, so the restrictions that apply to Sarah cannot take hold via that route. Diana's respecting of Ernie's CA means that she cannot instil any unsheddable values to ensure *A* and *E*. She cannot depend on a deliberative failure on Ernie's part, or on his being coerced into *A*-ing; and she must leave Ernie open to rational persuasion (in a suitably CA-respecting fashion) But none of this implies that Ernie needs to *be* rationally persuaded that he ought to *A* thus bringing it about that *E*.

There are possible scenarios in which Diana (intentionally) achieves her ends by having it turn out that Ernie is rationally persuaded to *A* such that *E*. Perhaps there might be some conceptual value in comparing and contrasting this subset of Ernie scenarios with the case of Percy. But acknowledging this possible subset of cases involving Ernie should not obscure the bigger issue as to whether the analogy between the Ernie and Percy cases is valid *in general*.

While Mele (2006) does not consider any such analogy (or the compatibilist response to the zygote argument that motivates it), he does mention variations on the Ernie case that would suggest it might not be analogous to the rational persuasion case. Mele (2006, p193) speculates that giving different specifications of the content of the story about Ernie might provoke different intuitions about the case. He thinks, for example, that if *E* were the death of Ernie's aunt, and *A* Ernie's poisoning of her in order to inherit money that would relieve him of his financial troubles, then some reflective agnostics might think that Ernie was not blameworthy for his act – after all, Diana engineered his zygote with this event in mind all those years ago. And because they may judge that he is not morally blameworthy, they might

[101] For example, it is not clear that Sarah could intend something very different in *F* coming to pass to what she must persuade Percy to intend in *B*-ing without engaging in some form of CA-threatening deception.

further judge that his action was not free (if it had been free, this would suggest at least some degree of responsibility and blameworthiness).[102]

No matter the value and accuracy of these speculations, what is relevant to my purpose is the fact that Mele does not seem to think that an ethically problematic action like poisoning his aunt is an unsuitable candidate for Ernie's *A*-ing, whatever the CA-respecting restrictions that must be in place *ex hypothesi*. But then it is hard to see how the cases could be relevantly analogous, or analogous in a way that is helpful to the compatibilist, because it is hard to imagine how Sarah could engage in a suitable instance of CA-respecting rational persuasion that both (a) convinces Percy to perform an ethically problematic act without (b) our judging that Sarah's persuasive efforts were, in some ethics and freedom relevant sense, manipulative. Imagining Sarah rationally persuading Percy to poison his aunt while all the time respecting his CA sounds more like something out of Silence of the Lambs (i.e. Lecter persuading a fellow prisoner to swallow his own tongue) than a case that could help compatibilism by convincing us that Ernie is not being manipulated.

We should not, however, just take Mele at his word in thinking that *A* and *E* could fall within the realm of the ethically problematic (not to mention frivolous, trivial, etc.) where the prospects for CA-respecting rational persuasion do not look good. We should revisit Mele's own description of the original zygote argument scenario, in which Diana engineered things such that *Z*:

> …will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to *A* and will *A* on the basis of that judgment, thereby bringing about *E*. If this agent, Ernie, has any unsheddable values at the time, they play no role in motivating his *A*-ing. Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly exercises his powers of self-control and has no relevant compelled or coercively produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator. So he satisfies a version of [Mele's] proposed compatibilist sufficient conditions for having freely *A*-ed. (Mele, 2006, p188)

Ernie needs to rationally deliberate to the conclusion that it is best for him to *A*. But because he does not need to be rationally persuaded by some third party that it is best for him to *A*, we need to be able to imagine a much wider range of possible interests, values and projects that Ernie might have taken on, in a manner consistent with his CA, and that might lead to a broad

[102] Mele (2006) speculates further about the intuitive consequences of imagining *E* as a homeless shelter's receipt of a significant donation; and he speculates that some agnostics might reach different intuitive judgements in the case of Ernie killing his aunt. As implied in the text that follows, the details and implications of all these speculations are not relevant to my argument.

range of possible actions that do not, in themselves, sound like good candidates for non-manipulative, CA-respecting rational persuasion.

Perhaps a suitably creative compatibilist might try to persuade us that, given the combination of Ernie's characteristics specified above (mentally healthy, ideally self-controlled, a reliable deliberator, the absence of any compelled or coercively produced attitudes, etc.), Mele is wrong to think that Ernie could reason to the conclusion that it is best for him to poison his aunt.

Our defender of compatibilism might have two things in mind here. On the one hand, perhaps we should be more comfortable, and less apologetic, about the idea that what is rational and what is good or right generally coincide. From this perspective, it might then seem that Diana is subject to much the same restrictions in relation to Ernie, despite not engaging him directly in a process of CA-respecting rational persuasion, as Sarah is in relation to Percy. Perhaps, then, because each available candidate for *A* and *E* will range from ethically innocuous to virtuous, a number of agnostics might come to agree with the compatibilist that Diana's activities *with respect to Ernie's action A* are not any more obviously manipulative than Sarah's attempts at rational persuasion.

On the other hand, the compatibilist might want to propose a shift in emphasis in the analogy between Diana/ Ernie and Sarah/ Percy. Instead of the emphasis being on rational persuasion by a third party, the compatibilist may have had in mind an emphasis on Ernie needing to be *led* or *persuaded by reason* that it is best for him to *A*. We should, in effect, be making a three-way comparison between (i) an agent having being persuaded by reason that it is best for them to act in a certain way, in the absence of the intervention or interference, proximal or distal, of any other agent; (ii) an agent (like Percy) concluding that it is best for them to act in a certain way having been engaged in a process of rational persuasion by another agent (like Sarah); and (iii) an agent (like Ernie) who is persuaded by reason that it is best for them to act in a certain way, where the circumstances of them reasoning to this decision have been arranged at a distance by a third party (like Diana). The proposal is that (iii) closely resembles (i), given the proximal absence of any third party, while the CA-respecting constraints on (iii) would also make it closely resemble (ii). On this three-way comparison, it is thus proposed that Ernie reasons his way autonomously to his conclusion that it is best for

him to *A*, while Diana's arrangement of this state of affairs is no more nefarious nor manipulative than Sarah's presentation of the recommendations of reason to Percy (as in (ii)).

There seems little reason to concede anything significant to the first of these responses. Ernie's killing his aunt may seem extreme, but there are many other distasteful candidates for *A* that Ernie might reason it is best to do, given the myriad possibilities left open within the zygote argument for the projects and (sheddable) values that Ernie could come to have. He might refuse help to a stranger, invest money in an environmentally problematic enterprise, or vote for a conservative, benefit-cutting, anti-immigration politician in an election. And while we might respect the rights of autonomous agents to freely reason their way to such decisions in a liberal society, agnostics are unlikely to see Diana's arrangement of Ernie's future to ensure such distasteful outcomes as anything less than manipulation, whatever his proximal processes of CA-respecting deliberation.

The second part to the response suffers from a similar difficulty. Given a suitably rosy or innocuous candidate for *A*, Ernie's reasoning towards his decision to *A* under (iii) might have all the surface appearance of the independently deliberating agent under (i), with Diana's interference rendered apparently harmless by her need (desire) to respect Ernie's CA. But only slightly less savoury candidates for *A* help scratch through this appearance and return Diana's interference to a once-again clear case of manipulation. Ernie voting for a right-wing candidate because reason persuaded him to do so is unfortunate; his having voted this way because Diana arranged it such that he would be led by reason to do so is manipulative[103]. If our compatibilist is not bothered by the ministrations of Diana as the distal origin of Ernie's regrettable political decision, this is (again) their choice, given the positions on offer to a pre-commited non-agnostic.

But let us suppose, for a moment, that our interlocutor had persuaded us of some pertinent similarity between cases (i), (ii) and (iii), such that Diana is constrained in her choice of *A* and the means by which Ernie reasons to the conclusion that he should *A* in ways that might,

[103] Given enough time and enough zygotes (she is superhuman, after all), Diana might arrange a regime change with all the hallmarks of a democratic change of government freely brought about by CA citizens. That such a regime change is CA-respecting rather than based on brainwashing and/or indoctrination is definitely a source of potential interest; but it remains *Diana's regime change*, brought about by her manipulations of the zygotes of her pawn (future) citizens.

for a moment, remove the appearance of manipulation from the scenario[104]. There remains a gap in the zygote argument scenario that can be exploited to reveal the residual disanalogy between cases such as Ernie's and Percy's.

Recall that I suggested Sarah's intentions in having Percy bring about *F* would need to be consistent with Percy's intentions in *B*-ing, otherwise she would need to engage in some kind of potentially CA-threatening deception as to the point of Percy *B*-ing. What should be clear is that no such restriction applies to Diana. There is nothing in Mele's setting up of the zygote argument scenario that implies any specific relation between *A* and *E*, except that Ernie's *A*-ing is crucially involved in bringing it about that *E*. There may well be some kind of constitutive relation between *A* and *E*. Nevertheless, there is no requirement that the aspect or description under which Ernie intends his *A*-ing be identical or even consistent with the aspect or description under which Diana desires (and intends) *E* to occur[105]. Similarly, Ernie's *A*-ing may have a range of consequences that he does not intend, while these may well be included amongst Diana's intended outcomes of having Ernie *A*. Percy's *B*-ing might also have unintended consequences, but these could presumably not be included amongst Sarah's intentions in having Percy *B*, on pain of her deliberately deceiving or otherwise withholding information from him in a potentially CA-threatening manner.

Note that it cannot be objected that Ernie would also be CA-compromised if he were in some way deceived about the consequences of his *A*-ing, including the nature of *E*. No one need engage Ernie in a discussion about, let alone rational persuasion over, his *A*-ing. Certainly, as we have already highlighted, Diana engages in no such process. Diana arranges atoms in *Z*. Nothing that she does at this level could be mistaken for CA-threatening deception. *Contra* our compatibilist interlocutor, Diana is a manipulator, irrespective of whether or not her intentions are benevolent or malevolent. What she is not is a deceiver.

Diana is a superhuman predictor. She has made a perfectly reliable prediction, relative to the state and laws of the universe, plus her arrangement of atoms in *Z*, about how adult Ernie and

[104] I think that this would be to concede too much to the compatibilist.

[105] It might be suggested that, under these circumstances, Ernie would only be responsible for his action under the description he intended; but it is not clear how this helps the compatibilist case. Those inclined to see Ernie as a victim of manipulation, despite his apparent CA, will think he is less than fully responsible for his actions. On this view, the key problem for compatibilism is that it fails to pick out the relevant sense in which Ernie's agency is compromised. Highlighting a sense in which Ernie retains responsibility for his action does not help answer this charge.

his CA deliberations and actions will fit into a larger framework of events in 30 years time. So long as Ernie is, at worst, only ignorant (in a CA-consistent way) of some of the consequences of his *A*-ing, Diana can be as nasty as she likes in terms of the nature and consequences of *E*. That is, as long as there is nothing that Ernie ought reasonably to have known or foreseen, then his CA cannot have been threatened. It is unnecessary to put flesh on all the possible scenarios that might satisfy this malevolent schema, but one quick example might be useful. We might suppose, for example, that Ernie's *A*-ing might be driving somewhere at a particular time along a particular route, such that Diana's desired event *E* (of which Ernie's *A*-ing is partly constitutive) is the tragic and traumatic death of a child running out into the path of Ernie's oncoming car, in pursuit of their tennis ball.

In summary, even if we do not follow Mele in allowing that Ernie might knowingly poison his aunt, there is still ample room for Diana to capriciously toy with what *E,* and the consequences of *E,* turn out to be, whatever Ernie's intentions in *A*-ing. At the same time, we should remember that the content we give to the example to stand in for *A* and *E* shouldn't really matter. Diana is, and always was, a manipulator. She is a manipulator who, for whatever reason, respects her pawn's CA, but that does not make her any less of a manipulator. Nor would some benevolent outcome of her interventions make her less of a manipulator. Given her predictive powers, her intentions, and a deterministic universe, she has manipulated *Z* in such a way that Ernie's autonomy, and with it his freedom, are compromised. The claim that her activities are relevantly analogous to processes of rational persuasion cannot be sustained.

There is one further disanalogy between the cases that should dissuade us from thinking of Diana's manipulations along the lines of the influence of a rational interlocutor. Mele (2006) discusses what he regards as a slightly modified version of the original zygote argument scenario. The point of this slightly modified version is to highlight the sense in which Diana, given her predictive powers and efforts at precision engineering in *Z,* can become complicit in and responsible for how *everything* in Ernie's life turns out:

> …her means of achieving her aim [by arranging the atoms in *Z*] is a cause of all Ernie's actions and not merely of his *A*-ing… In [this] modified version of the story, Diana has a much more extensive aim – to create an agent who performs *all* of [Ernie's] actions. (Mele, 2006, p190; italics in original)

What has changed, in this slightly modified version, are Diana's intentions rather than her interventions. Now she intends for everything that happens in Ernie's life to happen, and to

happen because she arranged *Z* in the way that she did. And all of it will happen, as before, without any compromise to Ernie's CA.

But now the disanalogy between the cases is made all the more stark. We cannot for even a moment imagine how Diana's intentions in being a cause of *all* Ernie's actions could be considered analogous to cases of CA-respecting rational persuasion. No ordinary human life can be imagined where every action is somehow the (partial) outcome of a process of rational persuasion involving a third party, let alone the same third party. Likewise, it is difficult (and, in the light of Chapter 9 below, unrealistic and undesirable) to imagine a human life where an agent is led to their every action by the dictates of reason. But then, for some of us, it was clear at the outset that Diana would be a freedom-undermining manipulator, no matter how narrow or broad (or benevolent) her intentions.

Before closing the discussion, however, I would like to argue that there are more 'realistic' variations on the theme of pre-natal manipulation that, in a suitable deterministic universe of our imagining, would not require superhuman feats of prediction and atomic rearrangement. In the slightly modified version of the zygote argument we have just considered, we were asked to focus on the sense in which, having arranged the atoms on *Z*, Diana effectively becomes a cause of all that Ernie does. This idea of some general and pervasive effect of some pre-natal event/s suggests the possibility of some less fanciful[106] CA-preserving interventions that require some accurate predictive knowledge, at least of a suitable statistical kind, and yet which (unlike so many crudely imagined cases of genetic manipulation) do not require some form of rigid and irreversible hard-wiring.

Suppose it was discovered by scientists that a particular substance – let's call it, for reasons that will become apparent, 'hypoaesthica' or HTA for short – is reliably correlated with low levels of artistic interest and ambition. Specifically, levels of artistic interest and activity in adulthood, including engaging in artistic hobbies, pass-times and careers, are negatively correlated with exposure to HTA in the womb. At the same time, it is discovered that the effects of HTA exposure in adulthood are neither inescapable nor irreversible: adults exposed

[106] Well, at least, considerably less fanciful than the scenario imagined in the zygote argument. I have tried to make the example that follows vaguely amusing, but not so as to highlight or exaggerate its implausibility. Pre-natal 'manipulation' is arguably not a fanciful notion at all, as witnessed by the lengths to which modern mothers can go in both avoiding (e.g. drugs, smoking, alcohol) and adding (e.g. music, Mozart, dietary supplements, etc.) elements that have potential effects on the pre-natal environment.

to high levels of HTA in the womb reliably develop and exhibit a pattern of interests and values that disincline them from artistic activities and pursuits, but these interests and values are, in a suitable (compatibilist) sense, sheddable[107].

Let us further suppose that HTA is found to occur in high concentrations in Brussels sprouts, such that consumption of these sprouts during pregnancy will reliably cause exposure to excessively high levels of HTA for the foetus.

Publication of this research leads mothers around the world to avoid Brussels sprouts during pregnancy (to the extent that they needed any encouragement to do so). Unfortunately, the published results had a rather different impact on a close-knit community of puritanical farmers with a long-established dislike and distrust of all things artistic. Music, theatre, films, fiction, poetry, painting, sculpture: to members of this community, all these were worthless (not to mention morality-threatening) distractions from the hard toil required to make a success of life. Empowered by the discovery of HTA, however, the community embarks on an active process of increasing Brussels sprouts consumption during pregnancy in order to begin the process of effectively reducing, if not eliminating, the numbers of wayward artistic bohemians and layabouts in their ranks.

In defence of this otherwise unfortunate policy, the community members are quick to highlight that their policy of intervention will completely respect the autonomy and freedom of choice of the next generation/s, given that the effects of HTA exposure are neither inevitable nor irreversible. Moreover, the community will maintain the same opportunities for exposure to artistic works and pursuits as were previously available in the community, and no child will in any way be barred from artistic interests or pursuits. Art will still feature in school syllabi; and there will be no closing of theatres, galleries, concert halls and music stores. The fervent hope of the community is that this autonomy-respecting biological Puritanism will finally succeed where more socially-repressive and restrictive had forms failed in the attempt to eliminate the arts. In summary, the community hopes in time, over the course of perhaps a few generations, to be free of the scourge of frivolous aesthetic indulgence by way of the *free and responsible life choices of these new generations*.

[107] That is, in Mele's sense of 'sheddable'/ 'unsheddable' values, it is a practical psychological possibility for adults exposed to HTA *in utero* self-initiate a process of changing their interests and values regarding artistic pursuits.

I submit that we have, here, a more realistic variant of the kind of deck-stacking undertaken by Diana that is equally deliberate and equally undermining of autonomy and freedom, and yet which might equally be claimed to be consistent with respecting the CA of the new generations in this community. As was the case in the zygote argument, it is clear that the manipulation that has taken place cannot be equated with, or considered analogous to, a strategy of CA-respecting rational persuasion. (One might assume that this community of anti-aesthetes would have already tried various policies of rational persuasion, both CA respecting and not, over many previous generations, without success.) Pre-natal manipulation through Brussels sprout consumption represents a clear change in strategy from that of CA-respecting rational persuasion. And, as in the zygote argument, we are also being invited to share the intuition that this manipulative deck-stacking behaviour on the part of this puritanical community is not significantly different to the case where 'blind' natural forces achieve the same effect in a deterministic universe (Premise 2), while sharing the intuition that children of the next generation in this community are not free agents, and are not responsible, for all of their choices, especially in so far as these relate to artistic tastes and pursuits (Premise 1).

Variations on this argument can be developed, including ones in which we vary whether or not the community is open or secretive with the new generations about the adoption and goals of the policy. Perhaps these variations might be imagined to have effects on the efficacy of the policy. What is not obvious is that such variations (including a policy of secrecy) could have potential relevance to the CA of the new generations, given the limited extent of the intervention (i.e. on the diet of pregnant mothers). More important, however, is for us to note that if we thought that openness (or secrecy) about the policy might have consequences for its efficacy, and if we thought this because we thought the policy might somehow create resentment in the new generation, this would only lend weight to the argument that this must be considered a freedom and autonomy threatening intervention, whatever the CA status of the new generation[108].

[108] We might imagine, say, an indignant teenager berating their parents: "Well, of course I don't have much interest in the arts, as things have turned out; but you might at least have given me a fair chance to acquire those interests and make those choices, instead of stacking the deck for me before I'd even seen the light of day!"

In summary, it is my claim that the conclusion of the zygote argument is sound, and that the compatibilist attempt to deflect the force of the argument by suggesting an analogy between CA-respecting manipulation and CA-respecting rational persuasion fails[109]. The deliberate stacking of the deck before we are born involves manipulation that compatibilist accounts of autonomy struggle to recognise as such. At the same time, such deliberate stacking of the deck is not different, in kind, to the deck-stacking that naturally takes place in a deterministic universe before each of us is born. Both forms of deck-stacking undermine claims of freedom, however different things might look if we just focus on the proximal level and the important considerations compatibilists have highlighted there.

We thus find ourselves at risk of floundering in a stagnant and counterproductive impasse. On the more local and proximal level of deliberations, choices and action, libertarianism faces what looks like an insurmountable difficulty in trying to persuade us that somehow inserting indeterminism into moments of volition and action-production could ever establish and sustain claims of a stronger, special form of agency that we can claim to be free. And while compatibilism apparently fairs better at this more local, proximal level, it remains unconvincing at the global level when confronted with the potential implications of our living in a deterministic universe. Whatever their capacities for reflective endorsement, the sheddability of their values and motivations, and the conditional availability of alternative possibilities, agents in a deterministic universe find themselves causally straight-jacketed within a deterministic matrix of conditions not of their making, such that any given choice or action of a given agent could turn out, on closer examination, to be either 'theirs' or the clever manipulations of a Diana.

How we might move beyond this impasse towards an account of free agency that might satisfy our interest in, and secure our claims of freedom, while not simply changing the

[109] There will be compatibilists for whom the analogy is not critical. For the latter, what is crucial to thinking about and judging intuitions in Ernie's case is that Ernie must come to decide on his action by way of reason. This need not be strictly analogous to rational persuasion by a third party, only no more nor less manipulative than such persuasion. It is hard to see, however, why someone not already committed to sustaining a compatibilist position in the face of the zygote argument would take this view. Fill in the details of Ernie's CA-respecting reasoning, abstract these from the context of the zygote thought experiment, and it might look like Ernie had a compelling case for *A*-ing as he did. But the point of the argument is precisely to juxtapose this with Diana's existence as the manipulator who set all of this up, such that the coexistence of manipulation and compatibilist-style autonomy might expose the deficiencies of the latter when it comes to securing free agency. Reasserting that CA reasoning does the job nevertheless, if you are a compatibilist, doesn't really help move the case one way or another.

subject altogether relative to the issues and concerns associated with the traditional debate, will be the challenge to be taken up in Part II of the thesis.

Part II

Chapter 5

Changing the Subject without Changing the Subject: An Alternative Framework for Explicating Free Agency

The conclusion of Chapter 4 did not look promising for those interested in the defence of free agency: the incompatibilist suspicion is basically correct (you can't actually get free will in a deterministic universe), but so is the compatibilist (and hard determinist) suspicion about libertarianism (you can't actually get free will by adding indeterminism into acts of volition). This might be music to the ears of those who enjoy taking up the 'hard' positions in philosophical debates – the 'hard' or sceptical incompatibilism discussed in Chapter 1 certainly challenges 'folk' wisdom, philosophical orthodoxy, and our experience as agents. But to those who hoped to find the appropriate space within the parameters of the traditional debate to defend free agency, the news does not seem good. It would appear that the traditional debate really does run into something like Kant's Third Antimony of Pure Reason – a deep and seemingly irresolvable cognitive conflict between ways in which we view ourselves, our actions, and the way we and our actions fit into the world around us.

The fall-out from this frustrating impasse is, needless to say, considerable. For a start, there is very little apparent middle ground. Non-libertarians think that libertarians suffer from some kind of wilful cognitive failure by refusing to admit that indeterministic agency is dead-end; and incompatibilists think compatibilists suffer from some kind of wilful cognitive failure for refusing to admit that defending free will in a deterministic universe is a dead-end. But this is not the real problem.

The real problem[110] is that if you agree that there is real impasse here, and yet despite this you think that there is value in exploring and defending a meaningful account of free agency, it is extremely difficult to negotiate the necessary terrain without getting drawn into answering questions like: "How would what you are proposing/asking/debating/etc. help adjudicate the issues between compatibilism and libertarianism?" The response I have given to this question is that, in an important sense, the issue cannot be adjudicated much beyond

[110] Or at least, the one that is most stifling within the traditional debate.

what has already been said. Both positions have fundamental flaws[111]. What we need is to find a way to change the subject without changing the subject. That is, what we need to find is a way to say something constructive about, and in defence of, free agency where the issues deemed worthy of attention are no longer defined[112] by the landscape of possible positions and points of contention that are so well mapped out within the traditional debate.

The primary challenge of this chapter will thus be to lay out such an alternative landscape of issues and points of concern or contention within which it would be meaningful to locate an account of free agency that is not primarily preoccupied with questions about compatibilism, incompatibilism and libertarianism. I will argue that the task of elucidating and defending free agency can be constructively relocated within a framework in which two images help define the terrain: the image of an *agent automaton*, on one hand; and, on the other, the image of a *hyper-reflective, hyper-rational agent*. Defenders of free agency are thus to be confronted with the challenge of articulating an account of human agency that avoids the excesses of claiming we are some kind of hyper-reflective, hyper-rational agents while successfully resisting attempts to reduce us to the status of agent automatons.

However, before proceeding to expand on this alternative framework outlined by the images of the Agent Automaton (AA) and the Hyper-reflective Hyper-rational Agent (HHA), one final digression into the territory of the traditional debate is required in order to address, and postpone to some extent, a likely question given the incompatibilism implied by my arguments of Chapter 4: if I am espousing an incompatibilist view in which I wish to defend a claim that we are free agents while also rejecting traditional libertarianism, then how exactly do I propose to deal with the second horn of the dilemma (posed by Lipton (2004), as outlined in Chapter 1) in such a way as to make free will compatible with *indeterminism*? At the same time, given my arguments of Chapter 2, 3 and 4, it might be asked how I think I am able to avoid taking up a sceptical incompatibilism? Unproductive impasse or not, adherents

[111] Perhaps libertarianism looks more critically flawed than does compatibilism, because its errors seem more obviously internal to its account of volition; but if the conclusion of the zygote argument is correct, then compatibilism is also critically flawed in ways that will ultimately, if sometimes less obviously, play themselves out in accounts of volition. Compatibilism will look for and settle on characteristics of the agent, their will and their actions that they think can sustain claims of freedom and responsibility in a deterministic universe. But the arguments and diagnoses of the previous chapter suggest that the apparent merit of such compatibilist projects will tend to fade from view once a sufficiently global view is taken in which global determinism holds sway.
[112] Or no longer exclusively defined.

to the parameters of the traditional debate are likely to want some kind of answer to this sort of question, and it is to this challenge that I now turn.

A Compatibilism Concerning Free Will and Indeterminism?

As intimated above, a special challenge that someone taking my position faces is that I have already rejected (by way of my arguments against Kane's account in Chapter 2) the traditional libertarian proposal of inserting indeterminism into moments (or occasional special moments) of volition as a means to securing claims of free will and responsibility. It would seem, therefore, that I am committed to some form of free will-indeterminism compatibility thesis, but the most well explored options (in the guise of agent- and event-causal libertarian accounts) are off limits. What, if anything, can I sensibly propose while also doing justice to my convictions (and arguments) that there are neglected topics and questions about free agency that deserve more urgent attention than any lingering puzzles arising from concerns within the traditional framework?

I will limit my responses to a more programmatic level, both because I wish to turn to these more pressing topics, but also because my responses point to potentially substantial projects for future research that cannot be sensibly pursued here. Furthermore, I cannot claim that any answer I offer to the question posed above will come ready-armed with adequate responses to the various challenges and objections I would expect to encounter, especially from within compatibilist and hard incompatibilist quarters. With these qualifications and apologies in place, I will proceed to address the question at two different levels, and then point to one particular positive proposal whose substance and promise I will revisit in the penultimate chapter of the thesis.

The first level at which the question about indeterminism can be addressed involves a return to the conclusion of the zygote argument. The conclusion of this argument, like most other arguments against the compatibility of free will and determinism, was that there can be no free will in a deterministic universe because, compatibilism notwithstanding, Diana could manipulate an agent like Ernie in a deterministic universe.

If the universe is *not* deterministic there may or may not be free will, as far as the considerations of the zygote argument go. It would seem that the manipulations we imagined Diana capable of in a deterministic universe would not be possible in an indeterministic one –

at least, the combination of probabilistic causal relations with any significant degree of chaotic dynamics, especially in the developing brain of our young Ernie, are likely to render Diana's ministrations futile[113]. But there seems to be no specific implication of the zygote argument that some kind of indeterministic process/es *inside* agents is/are necessary for free will. It is enough that free will is no longer threatened by the truth of (global) determinism.

So, at this level of response, it is proposed that it is enough (a) that our universe is not deterministic, and (b) that we recognise that compatibilist attempts to ratchet or gear up the demands for free and autonomous agency in the face of deterministic threats (such as Diana's manipulations) need not be necessary in order to secure freedom of the will. If we no longer need to accommodate ourselves to determinism, incompatibilist conditions for claiming free agency need not be of the form, nor as necessarily demanding, as those proposed by the likes of (especially autonomy-based) compatibilist accounts. Most important, however, is that on this line of response, no straightforward inference is to be drawn from the rejection of determinism and compatibilism to the claim that indeterminism *must* therefore form a central part of the story we will tell about free agency.

Although I think there is an important point being highlighted above, I suspect that this first level of response may well be deemed insufficient by many within the tradition. A second level of response acknowledges the need to say something more, but situates such a task within a much larger project. What I have in mind here is the idea that the rejection of determinism (and, with it, the primary motivation behind compatibilism) is part of a much bigger project of rejecting and expunging reductionistic physicalist thinking from (especially, but not exclusively) the philosophy of mind and psychology[114].

Consider, once again, the Consequence Argument laid out in Chapter 1. One underlying idea driving the argument is that if you fix the initial physical state of the universe and the laws of nature, everything else thereafter – including events involving agents, their choices and actions – unfolds like clockwork. Rejecting determinism, or more specifically rejecting determinism by making the laws of nature ineliminably probabilistic, does very little to

[113] This is not to suggest that manipulation is not possible in an indeterministic universe – any manner of manipulations remain possible, including grandiose versions such as the scenario depicted in the film *The Truman Show*.

[114] Reductionistic physicalism is to be expunged, not in favour of some variety of dualism, but rather in favour of what Crane and Mellor (1995) once called 'egalitarian pluralism.'

change this image underpinning the argument – namely, the idea that the events unfolding at the level described by physics somehow fix and carry along everything else including, eventually, events involving and facts about us, our psychology and our agency[115].

Of course, recognising a pervasive role for indeterminism might help advance this larger project – for example, it might help by displacing a stubborn, Newtonian conception of billiard-ball-like causation[116], as well as destabilise various problematic conceptions of the laws of nature[117]. But, on their own, such results would not be sufficient for the advancement of the larger project within the philosophy of mind of defending the causal and explanatory relevance of the mental in ways that secure, for example, meaningful and defensible senses for concepts like 'downward causation', not to mention a clear and unambiguous causal role for consciousness. Gesturing towards this larger project, and the direction I think it ought to take, is obviously something of an unsatisfying promissory note response to the question posed about the role I foresee for indeterminism. But it should be clear that this larger project, if pursued here, would involve *several* projects of future research, and would delay indefinitely the task of making more immediate progress in understanding and defending free agency outside the parameters of the traditional debate.

Given the above responses, comments and promissory notes, it is not my intention to say too much more about the role/s I foresee for indeterminism in securing our claims of free agency. However, the discussions so far – in particular, my discussion and critique of Robert Kane's work in Chapter 2 – and the proposals I make in Chapter 10 for exploring the significance of imagination to free agency do allow for at least one speculative hypothesis about a possible place for indeterminism *within* the psychology of a free agent, and where such indeterminism does not threaten the integrity, ownership or control of the agent in the ways I have alleged for traditional libertarianism.

---

[115] Obviously, this is too quick – there is much more to be said to establish the link I have in mind; and there is nothing in the skeletal outline of the Consequence Argument that rules out, for example, the possibility that the laws of nature might include ineliminably psychological laws. Nevertheless, even on a more careful presentation, I would claim that the idea of fixing the values of physical variables (psychological variables not being obvious candidates for having values at the time of the Big Bang) remains the underlying driver of the reductionist physicalist image at work in this kind of argument.

[116] Recall Kane's (2002a) complaint, discussed in Chapter 2, that we are far too prone to equate 'undetermined' (in the sense of indeterministically caused) with 'uncaused.'

[117] I have in mind here, for example, the work of Nancy Cartwright (1983, 1999) and John Dupré (1993).

As I have noted a number of times, the central conclusion of my evaluation of Kane's account was that under traditional libertarianism, as represented by Kane, the attempt to insert indeterminism into moments of volition tends to undermine claims of ownership and control of action – the compatibilist suspicion about libertarianism is basically correct. At the same time, my discussion of Kane's work highlighted two important, positive features of his account. First, Kane proposed to help ground claims of Ultimate Responsibility by positing occasional, special moments of self-shaping – his SFAs – in which undetermined events within the agent might come to shape their future reasons and decisions, determined or not. Indeterminism need not be inserted into each and all instances of choice in order for these to be the choices of a free agent. It could be sufficient that, in free agents, we could in part trace their choices back to these special, undetermined moments of self-shaping. Second, particularly in the context of defending Kane against Dennett's (2003) criticisms, I allowed that it could matter, in principle, both that an indeterministic process of self-shaping occurred *inside* the agent[118], and that it could be significant that the indeterminacy involved was attributable to the psychology of the agent[119]. Is it possible to put variants of these ideas to work in defending an incompatibilist account of free will that avoids the libertarian tactic of inserting indeterminism into moments of choice and volition?

The speculative answer, which I will expand on briefly in Chapter 10, is that indeterminacy within certain processes involving *imagination* could be recognised as an important ground for our claim to being free agents. While I will delay offering more details until that later chapter, I will endeavour here to give a rough sketch of the hypothesis I have in mind.

As I argue in Chapter 10, imagination is a topic that appears almost entirely neglected within contemporary discussions of free will[120]. And yet, at an intuitive level, it seems obvious that imagination provides the space in which to generate, contemplate and test out the consequences of alternative possibilities for action. At the same time, the domain of the imagination is rich with associations of creativity, novelty, origination and ownership – concepts, as we have seen, that play an important role in various conceptions of free agency. My (less controversial, initial) proposal is that imagination should be accorded its rightful

[118] As compared to Dennett's (2003) idea that being connected, by remote, to a geiger counter would achieve the same ends in resolving the tension in a SFA.
[119] As compared to the anti-libertarian tradition, well-represented by Dennett (2003), of insisting that randomness is just randomness.
[120] I offer persuasive, if defeasible, literature-based evidence for this claim.

place amongst the capacities that ground our free agency. More speculatively, however, I want to suggest that in the domain of the imagination, indeterminism might find a place where it enhances and enriches our freedom, instead of in some way threatening or undermining it.

The core of this second, more speculative proposal derives from the idea that in the case of the imagination, and unlike processes such as reasoning, deliberation and choice, we might *welcome* the open-ended generativity and possibilities for novelty that could result from a degree of indeterminism *without* thinking that this probabilistic (or, if you prefer, chance) element within the process somehow undermined claims of origination, ownership and control on the part of the agent. That is, the open-ended generativity of imagination when conceived of as a partly indeterministic process could be seen as a virtue and an aid to free agency, not as some kind of shortcoming. Conversely, we might make a case for the freedom of an agent being restricted, curtailed or otherwise undermined when there has been a foreclosing on the generativity of various imaginative processes at play in not only contexts of decision and action, but also in processes of agent shaping or formation more generally.

In the realm of imagination, then, perhaps we may find a process (or processes) in free agents where the insertion of indeterminacy brings (i) the added values of generativity, novelty, an expansion of perspectives and of imagined possibilities, while (ii) the associated ceding of some degree of control[121] over such imaginative processes seems tolerable, indeed desirable, to the extent that (iii) attempts at reasserting too much control, thereby foreclosing on the unfolding imaginative activity, would tend to constrain or threaten freedom of agency.

As indicated above, this positive, speculative hypothesis concerning indeterminism and imagination will be explored further towards the end of the thesis. Clearly, to the extent that I would be comfortable for labels from the traditional debate to be applied to my position, my view is a brand of *non-sceptical* incompatibilism. Can I say anything more as to why, given my rejection of traditional libertarianism, I think I can and should avoid a *sceptical* form of incompatibilism? I offer two reasons for my optimism, and my continued pursuit of a positive account.

[121] Recall Kane's (2002a) concession, discussed in Chapter 2, that the indeterminacy posited in his SFAs does imply a relinquishing of some degree of control.

First, as intimated above, rejecting the tactic of inserting indeterminism into moments of volition does not imply choosing the second horn of Lipton's (2004) sceptical dilemma – indeed, I would not propose the hypothesis just discussed if I thought the possibilities for harnessing indeterminism to secure free agency had been exhausted. Indeterminism within the agent might still form an important ground for freedom (as per my hypothesis regarding imagination). It is also conceivable that indeterminism could play a different role altogether in accounting for free agency – for example, by providing the relevant background 'default' when it comes to causal processes, against which human capacities for reason-guided, goal-directed action over intermediate- and longer-term timeframes might be a notable, freedom-grounding exception[122]. Rejecting traditional libertarianism does not, therefore, imply embracing sceptical incompatibilism about both determinism and indeterminism.

Second, the arguments of Chapter 4 imply that there is no free will to be had in a deterministic universe. In this sense, compatibilism fails in so far as it is a project of trying to secure free agency irrespective of whether or not (global) determinism is true. But as should be evident from Chapter 3, that may still leave much that is valuable in compatibilist accounts about the *local* conditions for and capacities involved in free agency. By the arguments of Chapter 4, I am committed to the claim that compatibilist conditions for freedom will not be sufficient in a deterministic universe: you can't secure freedom just by 'ratcheting up' the levels of rationality, reflective endorsement, self-control, etc. of agents, irrespective of the truth of determinism. By the arguments of Chapter 9 of Part II of the thesis, I also think that there are serious risks and problems that come with the strategy of trying to secure freedom via overly strong claims about our capacity for reflective, rational agency. But again, this does not mean that an appropriate incompatibilist account of free will won't emphasise many of the features of human agency rightly prized by various brands of compatibilism. Our capacities to follow and be responsive to reason; to give and demand reasons of ourselves and others; to deliberate along roughly rational lines; to reflect on, endorse and shape our own projects and values – these are important and valuable features of human agency that must feature in any adequate defence of freedom, especially once (following the evidence and arguments of Chapter 9) our claims to such capacities have been appropriately investigated and qualified in the light of empirical evidence. Thus, rejecting compatibilism *qua* an account

[122] A position along these lines is put forward by John Dupré – see especially Dupré (1993).

of free will that is compatible with the truth of (global) determinism does not leave traditional libertarianism, hard determinism and sceptical incompatibilism as the only alternatives.

Do these considerations warrant an outright rejection of sceptical incompatibilism? It would be arrogant, not to mention premature, to suggest that they do. My primary purpose at this point is far more modest – namely, to provide a rationale for continuing with a positive project aimed at defending free agency. As will soon become evident, I think that there are more pressing and productive puzzles and problems that require attention. Articulating and responding to these puzzles and problems should push us towards a more nuanced and empirically-informed account of human freedom that is better able to respond to the sceptical incompatibilist threat; but fully developing that account of freedom, and thus responding more decisively to the sceptical threat, must be work for a future project.

For now, I wish to return to the more pressing task of proposing an alternative framework within which to situate, explore and defend claims of free agency in ways that promise to generate new puzzles, motivate fresh perspectives, and call for different kinds of data and arguments to those that tend to dominate the traditional debate.

The Spectre of the Agent Automaton

Blackburn (1999) characterises the fear of determinism as mistakenly thinking that we only have two possible options when it comes to free will – either a ghost-like existence as a soul floating above the ebb-and-flow of material events (libertarianism), or something more like a tram relentlessly running on its single pair of rails towards its predetermined destination (compatibilism/hard determinism). There are two significant elements to the latter image. First, there is the image of a singular, predetermined path traced out by the rails. This aspect of the image is most obviously associated with the issue of alternative possibilities, about which compatibilists (and incompatibilists) have had much to say. Second, however, is a less vivid but equally troubling image or connotation of automaticity and inevitability – that a tram will, in some sense, trundle on regardless towards its destination, with the agent-self more in the role of passive passenger than active driver. Compatibilists (and incompatibilists) have had much to say about this image too, including (in the case of compatibilism) helpfully pointing out that determinism should not be mistaken for fatalism. Human agents, even in a deterministic universe, might still need to deliberate, choose and act – at least some of the time. As a Kantian might put it, human agents must act under the idea of freedom.

But the spectre of the agent automaton in this second image is not to be so straightforwardly dispelled. Towards the end of his book *Agents and Causes*, Timothy O'Connor (2000) expresses a concern that may have been felt by many who have delved into the traditional free will debate in search of insight into the nature of human agency:

> Something the philosopher ought to be able to provide some general light on is how consciousness figures into the equation. It is a remarkable feature of most accounts of free will that they give no essential role to conscious awareness. One has the impression that an intelligent automata (sic) could conceivably satisfy the conditions set by these accounts – something very counterintuitive. (O'Connor, 2000, p.122)

We cannot, of course, review all extant accounts of free will in order to properly evaluate the validity of O'Connor's claims. And we should note that, as an agent-causal libertarian, O'Connor no doubt has compatibilists in mind, for the most part, when making these complaints. There are, however, three elements here that might be helpful in trying to articulate what worries about agent automatons could amount to.

First, O'Connor finds it remarkable that consciousness is not generally given a prominent, nevermind an essential role within accounts of free will. It seems, as Searle might put it, that consciousness is just one of the most obvious and remarkable features of our waking lives as psychological subjects; and that conscious awareness of ourselves, our intentions and our actions is one of the most obvious and remarkable features of our lives as agents in the world. From a philosophically neutral point of view, the idea that we could either encounter or produce a satisfactory account of free will that did not give a prominent and essential role to conscious awareness is, *prima facie*, immensely puzzling.

Second, O'Connor sees it as a consequence of this failure to give proper place to consciousness, and conscious awareness, that many accounts of free will lay out conditions that could conceivably be satisfied by an intelligent automaton. This claim is likely to be far more controversial than the first, not least because different interpretations and connotations of the terms 'automaton' and 'intelligent automaton' are likely to elicit different judgements about the significance of O'Connor's claim, even if it is true. A sceptic about consciousness, and/or a sufficiently materialist compatibilist, might be happy to simply accept that human agents just are intelligent automata in the final analysis, in which case O'Connor has simply highlighted an aspect of their position with which they are in no way uncomfortable. For O'Connor, however, it would seem that an intelligent automaton is something more like a

mere mechanism lacking the conscious awareness that is so distinctive of human agents, and that should therefore feature prominently in accounts of human free agency.

Thus, as a third element to his complaint, O'Connor asserts that the idea of a *mere* mechanical system lacking in conscious awareness *also* satisfying a putative set of conditions for possessing free will is a deeply counterintuitive, and thus problematic, result. For O'Connor, automata lacking in conscious awareness just ought not to be candidates for possessing free will. This is an assertion rather than an argument, but the basis for the intuition seems clear enough. Something about the image of an automaton, intelligent or not, seems incompatible with what we might think is required to aspire to the status of a free agent.

I think we can fill out the complaint here, and perhaps strip it of any undue agent-causal libertarian motivation that O'Connor might have had in mind. We are conscious agents. Being conscious, and having conscious awareness of what we are doing during most of our waking lives, seems absolutely central to the kind of agents we are. To the extent that we are, or can, also be free agents, our conscious awareness seems equally central to our conception of ourselves as such free agents. We experience, exploit and appreciate our freedom, in significant part, by the involvement of our conscious awareness in our engagements as agents in the world.

There are, of course, numerous other features that we associate with our free agency. We can generate and make choices amongst options; we are often flexible and adaptable in our responses; we can learn and we can change; and we take ownership of what we decide and what we do. This list is not exhaustive, and it has not been constructed in a careful and carefully qualified manner. It captures much that philosophers have tried to explore within the context of the traditional debate about free will, without drawing the noted features into a coherent and detailed account.

Now let us return to the issues of consciousness, conscious awareness, and intelligent automata. We are able to design and build systems of increasing complexity and intelligence that could plausibly be claimed to exhibit many of the characteristics just listed: generating options, choosing amongst them, doing so in ways that are more or less flexible and adaptable, especially in so far as the systems are able to learn from past situations and

responses. But many of us are reluctant to think that just combining these kinds of features in an artefact of sufficient functional complexity would, as it were, suddenly give us a system that qualifies as a free agent. And a crucial motivator of this reluctance is the belief that intelligent and flexible behaviour in the absence of conscious awareness and control cannot amount to free agency. Conscious awareness and conscious control matter in crucial ways to the kinds of agents we are, such that systems that lack such awareness and control are not candidates, or good models, for attributing, understanding, or defending free agency.

These are substantive and potentially controversial claims about the significance of consciousness to human agency in general, and to the project of understanding and defending claims of our being free agents in particular. Whether or not all of these claims can be sustained is a matter that requires appropriate philosophical and empirical inquiry. For the moment, however, it is worth noting that there is nothing said so far about consciousness, conscious awareness and free agency that is obviously question-begging against one or other camp within the traditional debate (even if, as I have speculated, O'Connor might have mostly had compatibilists in mind when making his version of the complaint), mainly because nothing has been said or presupposed about the compatibility or otherwise of an essentially conscious free agency with determinism. There is nothing to stop a committed compatibilist being just as worried about the significance and role of consciousness as a libertarian might be. It is worth reflecting on why this might be so.

The spectre of a non-conscious intelligent automaton presents at least two difficulties that should make us reluctant to model free agency on the activities and capacities of such systems. The difficulties I have in mind both relate to self-awareness and related self-knowledge or understanding. The difficulties can be usefully presented as a kind of dilemma. On one horn, the functional capacities of a sufficiently complex and adaptive non-conscious automaton are taken to show that consciousness is not essential to the kind of 'self-awareness' and 'self-knowledge' required for free agency. Sufficient monitoring and feedback between systems, subsystems, and/or hierarchies of subsystems, all of them equally non-conscious, is claimed to deliver all the requisite system-level 'knowledge' required for a form of agency that 'knowingly' engages in sustained and coordinated 'intentional' activity of a kind with that undertaken by free agents like ourselves. On this horn, it turns out that we are not self-conscious and self-aware in the special ways we might have thought we were. This would represent something like the triumph of a Dennettian view of the conscious agent

(*a la* Dennett, 1991), brought down to earth with a thump of recognition that we are nothing more than just such a sophisticated combination of non-conscious sub-personal systems and subsystems (Dennett's "demons"), and yet still entitled to claim the title 'free agent'.

On the other horn of the dilemma, it is conceded that we have a form of extended or higher-order consciousness that is distinctive, and that brings with it a self-consciousness and self-awareness that does indeed enable equally distinctive forms of self-knowledge. On this horn, the success of modelling our agency in non-conscious intelligent automata has as its cost that this distinctive extended or higher-order consciousness is rendered impotent and epiphenomenal since, *ex hypothesi*, the non-conscious automata lack consciousness while successfully reproducing a relevantly similar form of free agency. Moreover, given that our experience as self-aware agents contradicts this epiphenomenal impotence, it turns out that much of our apparent self-knowledge must in fact be illusory. We have a distinctive capacity to know ourselves as self-aware agents, but we somehow get ourselves wrong, at least when it comes to the significance and impact of this self-awareness.

On this view, then, worrying about the threat of our being agent automatons (AAs) involves at least two possible concerns. Either we are going to turn out to be the sort of agent that rumbles along on some kind of un- or non-conscious autopilot, because it turns out that we lack the kind of extended consciousness and self-awareness we thought we had; or we turn out to be the sort of agent who (also) runs along on un- or non-conscious autopilot while experiencing, via our capacity for extended consciousness and self-awareness, an illusion of ourselves as being the conscious governors of our lives.

Having identified the threat, and having noted that it need not beg any questions against either libertarians or compatibilists[123], it should be noted that there is a more direct route to raising the prospect that we might be, or be very much like, AAs (agent automatons). This direct route need not have anything to say about the prospects for an actual automaton satisfying the conditions for free agency. Instead, the direct route to the threat of our being AAs runs through direct challenges to what we believe about consciousness and conscious agency. One such route would be through the kind of scepticism about consciousness evident in Dennett's (1991) *Consciousness Explained*. But another route, with more explicit links to

[123] Except in so far as a certain kind of compatibilist, such as Dennett, might claim that there is no problem, and that they are perfectly happy with the first horn of the dilemma.

questions about agency and freedom, runs more directly through empirical work on consciousness, agency, and our experience of agency. As we will see in greater detail in Chapter 6, empirical work on automaticity, on the apparent timing of conscious intentions in action, and on the experience of conscious willing, can and has been used to advance sceptical conclusions about the significance and role of consciousness in the initiation and control of human behaviour.

In advance of the details, we can note the gist of this evidence and the claims about agency, consciousness and the will that have been mounted on them. The bulk of the evidence can be roughly divided into two. On one hand, there is a range of evidence taken to suggest that we have far less conscious control over our mental lives, our choices and our behaviour than could be consistent with any strong claim that we are free agents[124]. There is evidence, for example, that we can have various cognitive goals, behavioural goals, and evaluations of persons and objects activated outside of our awareness with significant consequences for our behaviour – behaviour that thus appears to be automated in a significant sense[125]. And more sceptical interpretations of Libet's famous studies on the timing of conscious intentions[126] would suggest not only that conscious intentions lag behind unconscious neural processes that represent the 'real' initiation of behaviour, but also that the picture of action initiation that emerges from these laboratory studies leaves little (conceptual) space and (even less) time for consciousness to make a difference to what we do.

On the other hand, varieties of evidence have been highlighted that suggest we are regularly (and thus, potentially, are in general) wrong in our own claims about, and interpretations of, our motivations, intentions, and conscious involvement in the initiation and control of behaviour. So, in the most detailed version of this 'complaint', Daniel Wegner[127] strings together examples of mistaken first-person causal attributions as evidence for his overarching claim that our experience of conscious will – of our conscious thoughts being part of the causal chain leading to action – is *always* illusory.

---

[124] See contributions to the July 1999 issue of *American Psychologist*, including Bargh and Chartrand (1999) and Wegner & Wheatley, 1999.

[125] See, for example, Bargh and Chartrand (1999), Bargh (2008), and the various studies by Bargh and colleagues whose results are discussed in Chapter 6. Wegner (2002) also cites and makes use of automaticity-related results in his work.

[126] See especially Libet *et al.* (1983) and Libet (1985, 1999, 2004). Libet's data are discussed in both Chapter 6 and Chapter 7 below.

[127] See, for example, Wegner (2002) and Wegner & Wheatley, 1999

Taken together, this empirical evidence opens up a direct route to a sceptical view on which we are, as a matter of empirical fact, agent automatons who operate under various illusions as to the nature of our own agency. We do things we are not aware of, under the influence of factors and processes we are not aware of, such that when we get to explaining ourselves to ourselves and to others, we regularly (or even, to a significant extent, generally) get our motives, influences, and the springs of our actions[128] wrong.

We should not think, however, that the spectre of the AA is exclusively to be associated with issues of consciousness, self-awareness and self-knowledge, and the significance of these to human agency. Returning to Blackburn's (1999) image of the tram, we can interpret at least some of the threat posed by certain ideas about determinism as having a grip on us because of the threat of AA, rather than concerns about larger (metaphysical) questions about causation and the way the universe operates. The second image I associated with Blackburn's tram was that of the agent as passenger rather than as driver. This image does not require global determinism in order for it to trouble us. A certain amount of the 'wrong' kind/s of local determinism will do. If we turn out to be much more the irreversibly hard-wired products of our genetic inheritance than we think we could (or want to) be; if we turn out to be much more like the stimulus-response mechanisms Skinner and other radical behaviourists imagined us to be; if we turn out to be too much under the sway of hard-wired evolved and inherited traits, dispositions and modules; if we turn out to lack much of the flexibility, adaptability or plasticity that we think we have (and that some, like Blackburn (1999) see as central to any claim we might make to being free agents): all of these possibilities, rightly (but, perhaps, often wrongly) associated with various more deterministic approaches to human nature and behaviour, cohere into a potentially amorphous yet nevertheless tangible threat that we might be AAs with vanishingly little influence or control over the shape and direction taken by our life-paths.

In summary, the image of the AA is one of an agent whose activity in the world tends, on the whole, to be passive and responsive, largely automatic, unconscious, and instinctive. They might not lack consciousness, but their being conscious makes little apparent difference to their agency; and, if they are self-aware, they may suffer under various illusions as to the

[128] While Mele probably can't claim to have invented the phrase, my immediate source for "springs of our actions" is the title of Mele's (1992) book.

ways in which their being conscious matters. Their behaviour may or may not be deterministically governed at a local level, but to the extent that it is, it is governed by factors and processes that tend to bypass the 'will' (as a compatibilist might put it). They might be able to learn, but such learning is much more like the slow and relatively passive conditioning described by behaviourists. They are agents running along as if on autopilot. And they threaten our image, and intuitive models, of what a free agent is (and should be).

If the spectre of the AA is tangible and threatening in the ways I have suggested, why does it not feature more prominently in the traditional debate? We don't, presumably, want to construct a straw man (or a Dennettian 'bogeyman') just in order to generate movement in the wake of the impasse described in Part I. If we can't see the spectre of the AA obviously looming in, and shaping, the traditional debate, why might this be?

Part of the answer lies in the way in which the traditional debate, or the traditional dilemma, about free will is set up. Consider, once again, Peter Lipton's framing of the traditional problem space introduced in Chapter 1. According to Lipton (2004), as we saw, the traditional problem of free will can be set up as a dilemma:

> First, everything that happens in the world is either determined or not. Second, if everything is determined, there is no free will. For then every action would be fixed by earlier events, indeed events that took place before the actor was born. Third, if on the other hand not everything is determined, then there is no free will either. For in this case any given action is either determined, which is no good, or undetermined. But if what you do is undetermined then you are not controlling it, so it is not an exercise of free will… [C]onclusion: there is no free will. (Lipton, 2004, p.89)

The overarching goal of Lipton's (2004) paper is to consider whether or not the alleged spectre of genetic determinism poses any kind of novel and/or additional threat to free will. His conclusion in this paper is, essentially, that concerns about genetic determinism do not *add* anything significant to our concerns about free will that was not already implied by worries about *generic* determinism (as unpacked by way of the traditional dilemma).

In other words, Lipton's (2004) argument amounts to a series of attempts to unpack the possible consequences of genetic determinism (or, more neutrally, advances in genetic science) with a view to seeing whether any new or distinctive worries about free will might emerge; and his conclusion is that genetic determinism doesn't *add* any worries about free will that we can't already recognise from serious reflection on the problems associated with determinism in general. At best (or, if you like, at worst), advances and claims in genetics can

illuminate or make concrete some ways in which *generic* determinism might threaten free will, but no more than this:

> It can be deeply disturbing to be forced to face the ways in which determinism would make it true of all of our actions that we could not have done otherwise, and advances in genetic research may make it increasingly difficult for us to ignore this depressing fact. But even if the threat is thereby made vivid, it is not thereby made new. (Lipton, 2004, p.100)

If we generalise the argument, then any deterministic theory of human behaviour – a neural theory, a psychological theory, a sociological theory – could be used to make vivid the threat posed by determinism; but it would do so without posing any threat that was not already posed by generic determinism as unpacked in the traditional dilemma.

For the record, Lipton's (2004) conclusion is not sound. What is true, in the context of the traditional debate and the problem space structured by what he calls the traditional dilemma, is that a deterministic theory *qua* deterministic does not add anything to generic concerns about the compatibility of free will with determinism. But that is because the traditional dilemma is really just an argument for a generic sort of incompatibilism. From this perspective, it clearly does not matter what content you give to a deterministic account of human behaviour – determinism rules out free will.

A compatibilist will agree with Lipton that a deterministic theory's determinism is, in itself, not an issue of special concern, but for precisely the opposite reason (i.e. because a compatibilist does not see a deterministic theory *qua* deterministic as presenting any special problem for free will). But a sensible compatibilism will work precisely by making distinctions between different sorts of deterministic accounts, illuminating which do and which do not pose a threat to claims of free agency. And, presumably, any strongly genetic theory of human behaviour will, depending on its details, pose at least a *prima facie* threat to free agency in so far as human behaviour is found to spring from sources that bypass the will.

And yet, even if we can anticipate this response from compatibilism to Lipton's (2004) argument, we can see how the problem space of the traditional debate can discourage engagement with substantive threats to free agency such as that posed by the worry that we might be some kind of AAs because of the way, for example, our genes shape us and our behaviour. The incompatibilists will most likely see no distinct or novel threat, thus there is little argumentative gain to be had by compatibilists addressing these specific threats. So the overarching framework of entrenched positions and well-worked-out argumentative strategies

centred on issues of determinism, indeterminism, compatibilism and incompatibilism, discourages the articulation of worries that we might be AAs based on various *particular* (especially empirical) concerns.

There is another argument – this time grounded in compatibilism – that similarly discourages attention to particular concerns or threats that we might be AAs. Consider, again, the spectre of genetic determinism. From a compatibilist point of view, it can be argued that there is little point in addressing apparent threats to agency posed by genetic determinism *within the context of the traditional debate between compatibilism and libertarianism* because libertarian accounts will be equally threatened or undermined by the threat of genetic determinism. That is, if genetic advances paint a picture of the springs of behaviour that increasingly bypasses the psychology and the will of the agent, then libertarianism will be no better off, in principle, than compatibilism in dealing with this threat. So, once again, there is little to be gained, in terms of movement in the compatibilism-libertarianism debate, by articulating and responding to such a threat in any detail. So the threat will be set aside, and the spectre of AAs fades into the background of our concerns about free agency, to be replaced once again by concerns about determinism and the challenge of producing a convincing compatibilist account.

These kinds of (dismissive) tactics can be employed in response to the empirical challenges presented in the following chapter. Take, for example, Libet's data on the apparent timing of conscious intentions. On one interpretation of these data, what we are presented with is not a particular puzzle involving free will, but a general puzzle about mental causation: conscious will seems to lag behind (unconscious) neural activity in ways that suggest the conscious mental activity might be epiphenomenal. But then, so the argument goes, we have a problem of mental causation that is perfectly general, such that compatibilist and libertarian alike will have to find something to say about it in defence of human agency. In which case we may as well shelve the threat – both sides will have to deal with it in their own time, and because the threat is common, dealing with it will most likely not advance either side of the overarching debate. Similar arguments might be made for setting aside or ignoring Wegner's claims about the illusion of conscious will.

These are, I think, compelling arguments as to why the image of the agent automaton does not loom larger in the traditional debate, and why the loose collective of threats to agency

that I am grouping together under this image has not necessarily been identified and described on a regular basis as a source of concern about free agency. There are, however, additional reasons for the sidelining of many of these issues, especially those relating to consciousness and the role of occurrent mental activity in generating behaviour.

First, we should note that much of Western philosophy of mind, especially of the late twentieth century, has had a tendency to be, at best, unclear and, at worst, silent about the role(s) of not just consciousness but of occurrent mental activity in general. There is a tendency for philosophers of mind to carry out crucial debates about agency – most notably debates about mental causation, but also discussions about free will, autonomy, weakness of will, etc. – in a manner that adheres to a rudimentary belief-desire psychology, where our psychology is claimed to be efficacious to the extent that we are able to secure an ineliminable (essential?) role for reasons in causal accounts of our behaviour[129].

Whatever the value and plausibility of these efforts, reason-based accounts of action and agency are not straightforwardly open to being mapped onto occurrent mental activity. This is in large part because it is generally not clear how the usual components of reasons – propositional attitudes like beliefs, desires, pro-attitudes, etc. – are to be related to occurrent mental states associated with and (especially) preceding action. Moreover, arguments for holism about beliefs and desires, and for externalist accounts of the contents of such states, each represent trends in thinking about these candidate mental causes that, in various ways, discourage unambiguous pronouncements about the role of occurrent (not to mention conscious) mentation.

Both holism and externalism make it unclear how an occurrent mental state or activity, such as conscious thinking about a belief or desire, should be mapped onto claims (a) that an agent has a given belief or desire; and/or (b) that a given belief or desire played a causal role in generating a given action. Given holism[130], having a particular belief or desire depends on an agent having many other beliefs and desires, and not all of these (or even a potentially sufficient number of these) can be plausibly assumed ‘present’ or active in an instance of

[129] The classic statement of such a reason-based view of mental causation is, arguably, to be found in Davidson. See especially Davidson (1963).
[130] Davidson is, again, an obvious choice for authoritative source here – see Davidson (1970, 1974).

conscious thought about the particular propositional attitude. Given externalism[131], it is generally not clear that the content of a given belief or desire is always sufficiently transparent to the agent to allow a simple equation of consciously entertaining the thought of a belief/desire with the agent's correctly being attributed a belief/desire with a specified content.

The end result is that we, arguably, have available to us any number of reason-based accounts of mental causation that, even while they may try to reassure us of the causal efficacy of our minds and mental states by making reasons the causes of actions, they do not provide us with a very clear picture of how our mental *activity* contributes to our agency – most notably including our *conscious* mental activity. And while that might be a vaguely acceptable state of affairs when we limit our concern to overarching questions about the causal efficacy of the mental, it is hard not to sympathise with O'Connor's (2000) earlier complaint that this is unsatisfactory in the case of agency and free will.

When we ask questions about free agency, when we ask what difference *we* make in the world and what kinds of options and choices *we* have available to us, the 'we' that we most obviously have in mind here is most strongly identified with ourselves as conscious agents (rather than, say, merely agents for whom there are various true psychological descriptions or propositional attitude attributions) with varying degrees of awareness of what we are doing in the world. And this conscious awareness needs must be related to our ongoing activity as agents and subjects of experience. Whatever the issues at stake here, my immediate claim is that this state of affairs within the philosophy of mind does help explain the apparent absence of puzzles involving agency framed in terms of the threat of AA.

Mentioning consciousness again also suggests a further reason why we might expect consciousness to not play a prominent role in many extant accounts of free will. O'Connor himself offers a partial explanation, in the passage immediately following the extract quoted earlier: "That accounts of free will fail to provide an essential role for consciousness is nonetheless not surprising, given that its basic biological functions are presently quite mysterious to most theorists." (O'Connor, 2000, p.122). To the extent that O'Connor is correct in this pessimistic judgement of the state of consciousness studies, we might expect

[131] For important presentations of externalism about mental content, see Putnam (1975), Burge (1979) and Davidson (1987).

philosophers to think themselves ill-inclined *and* ill-advised to compound problems and puzzles about free agency by tying these too closely to the many and varied contentious issues associated with the study of consciousness – not least of which is the trouble that seems to arise when it comes to simply agreeing on a definition of consciousness.[132] By his own lights, then, O'Connor (2000) seems aware of the possibility that linking 'mysteries' of free will to 'mysteries' of consciousness may only succeed in making progress on questions about freedom problematically contingent on progress on questions of consciousness.

Linking this to our current concern, this apparent reluctance to link puzzles about free agency to puzzles about consciousness provides an additional reason for the spectre of the AA not to feature too prominently within the traditional debate. To the extent that concerns about our being AAs are linked to concerns about the significance and role of consciousness in agency, we should expect these concerns to have been downplayed or otherwise avoided as an unhelpful distraction within a problem space that is challenging enough without raising additional questions about consciousness.

So much, then, for an initial sketch of the spectre of the Agent Automaton, the loose but troubling collective of threats it poses to free agency, and the reasons we might not expect to find this image playing too prominent a role within the traditional debate. The sketch will continue to be filled out as we encounter specific challenges to free agency, most especially in the next chapter. I need to say more about the other image that is at play in, and that can help structure, both our worries about and our defences of free agency: the image of the hyper-rational, hyper-reflective agent (HHA).

Hyper-rational Hyper-reflective Agents

The image of an hyper-rational, hyper-reflective agent is one of an agent who is not necessarily a poor candidate for the status of free agent. Instead, the threat we can associate with the image of the HHA is that we might *not* be such agents and, *for this very reason*, we would fail to make the grade as free agents. That is, the 'treat' or risk involving HHAs is that we might mistake ourselves for and/or otherwise aspire to be such agents, thinking that our being or doing so would somehow secure our claims of freedom when, in fact, it generates a new set of problems. On one hand, there are potential problems involving whether or not we

[132] See, for example, Guzeldere (1995a, 1995b).

are or should want to be HHAs in the first place – i.e. questions about the normative desirability of certain characteristics of hyper-rational and hyper-reflective agency. And, on the other hand, there are potential problems of raising the bar for qualifying as a free agent to that of being an HHA, only to find out that we are not such not such agents. In general, the threat to be associated with the image of the HHA is that we might mischaracterise both free agency itself, and ourselves as candidate free agents, by linking freedom to a rarefied set of cognitive and behavioural attributes we might not possess, and that might not be all that desirable.

What, then, is a hyper-rational, hyper-reflective agent supposed to be like? Roughly speaking, HHAs are agents who tend to manifest (and/or aspire to) the following kinds of characteristics: they (i) exhibit maximal levels of self-control[133]; (ii) are logical, methodical and reliable in their reasoning and deliberation; (iii) act in line with their decisive best judgements about what to do; (iv) successfully constrain their wills over time, in line with these decisive best judgements; (v) are maximally self-reflective and self-aware; and (vi) manage to be affectively neutral, controlled and/or 'cool' in their reasoning and deliberation.

An immediate reaction one might expect to this list of characteristics, both individually and collectively, is puzzlement (or outright scepticism) as to why we would consider the image of such an agent as any kind of threat. Surely, it might be claimed, these are just the kinds of characteristics that we aspire to exhibit as agents – indeed, these are the kinds of characteristics alluded to by the likes of Dilman in Chapter 3, where we noted implicit developmental and normative claims built into the compatibilist idea of gaining ever greater insight, control and ownership over themselves as agents. Why should a move away from the traditional debate in any way undermine the apparent good (common) sense captured by such compatibilist ideas about reflective endorsement?

Such a challenge provides an opportunity for some immediate qualifications and clarifications about the image of the HHA I have in mind. It is in no way implied that the exercise of the capacities associated with each of these characteristics is never normatively desirable. Self-control, reliable and logical reasoning, identifying the best option for action and sustaining the effort to act in accordance with this judgement, self-insight, and not being

[133] Perhaps along the lines of an ideally self-controlled agent as described by Mele (1995, 2006).

a slave to one's emotional reactions: all of these have their place in the life of healthy, effective human agents. To argue that we are not HHAs is thus not to argue that we lack these capacities.

Instead, being wary of the image of the HHA involves questioning the extent to which it is both possible and desirable for human agents to instantiate maximal levels, or maximised versions, of these traits. Whether or not it is possible is ultimately an empirical question. Chapter 9 will involve confronting some relevant empirical considerations that bear on this issue, especially in relation to questions about the relationship between reason and emotion. Whether or not it is desirable is a much more complicated question. The arguments of Chapter 9 will also have some bearing on this matter. But we can say a little, in advance, on both fronts by revisiting material we encountered earlier.

In Chapter 3, it was noted that Mele (1995) characterised self control as follows:

> [S]elf-controlled individuals are agents possessed both of significant motivation to conduct themselves as they judge best and of a robust capacity to do what it takes so to conduct themselves in the face of (actual or anticipated) competing motivation. (Mele, 1995, p.5)

The philosophical literature on weakness of will is replete with examples of the many ways in which human agents might fail to fit this image of the self-controlled agent; the psychological and economic literature is equally full of examples apparently well-geared to bring our claims to self control down a notch or two[134]. So just what is possible, from an empirically-informed perspective, when it comes to feats and levels of self control amongst human agents is certainly open to further debate and inquiry. At the same time, it must be recognised that human agents do possess remarkable powers to take on projects and commitments that can require years of ongoing, self-sustaining effort and dedication. We are, in this sense, capable of truly remarkable feats of self control. But are we, and can we be, ideally self-controlled agents, given our many failures and shortcomings? It is far from clear that we are.

Is it normatively desirable that we should be self controlled, or ideally/ maximally self controlled, along the lines described by Mele (1995)? This is a much more complicated question. I cannot pretend to develop and defend an extensively worked out answer to this question here. But in service of my goal of defending the idea that HHAs pose something of a

[134] See, for example, George Ainslie's (2001) *Breakdown of Will*.

threat to free agency, I will make some brief critical remarks about the high regard in which self control appears to be held, especially amongst philosophers.

I take it that anyone advocating the normative desirability of ideal or maximal self-control must do so without helping themselves to any strong thesis of moral realism. There is no necessary connection between an agent judging that something is best for them to do, and that something being the ethically correct or desirable thing to do. This is both true for individual decisive best judgements, as well as for an agents' judgements in general. Given appropriate combinations of interests and motivations, self-controlled agents might judge that it is best for them to act in ethically problematic ways, whether in isolated instances (as most real-world agents probably do) or in general (as a sociopath might do).

Once we have separated out issues of self-control from questions of an agent's ethical values and conduct, I think it becomes far less obvious that ideal or maximal self control represents an attractive normative ideal for agents. Denuded of associations with doing what is right, Mele's ideally self-controlled agent might, under appropriate circumstances, be redescribed as a ruthlessly single-minded agent. And whereas pursuing a project of protecting the innocent with ruthless single-mindedness might, in general, be normatively desirable, following a blanket policy of obeying orders, or making money, or deceiving people, with ruthless single-mindedness would not. From this perspective, the self-controlled individual described by Mele might better describe any number of undesirable figures – the sociopath, the Nazi bureaucrat, the ideal soldier, the win-at-all-costs businessperson – whose lives, in various ways, lack balance, humanity, and a flexible sensitivity to the complexity and messiness of human existence. We correctly value a capacity for self control, but the value we attach to this is context bound and qualified. We do not value self control in itself, and we should not value or aspire to its ideal or maximisation. When we think about it, we know of (or can imagine) various maximally self-controlled individuals, and we don't really like them very much.

The argument for the threat posed by the image of the HHA is a generalisation of this kind of concern. Each of the features described under (i) to (vi) relates to important capacities of real human agents. Moreover, these capacities may, individually and collectively, have an important role to play in a realistic account of free agency. And yet for each of the characteristics mentioned, there may be good reasons to question whether or not it is possible

or (normatively) desirable for human agents to approach ideal or maximal versions of these traits. In short, we have various capacities that allow us to claim a degree of rationality and reflective self-awareness as agents, and it does not seem controversial that much of this is important to understanding and defending claims that we are, or can be, free agents. Yet it is not clear that we do approximate, or should aspire to approximate, idealised or maximised versions of these same traits that would render us some kind of hyper-rational and hyper-reflective super-agent. Thus the threat posed by the image of the HHA is that we should mistake ourselves for, and misguidedly aspire to be, an HHA when a characterisation and defence of our free agency requires no such distortion.

There is an additional line of argument, which I will develop in Chapter 10, for resisting the pull to make human agents look like HHAs. This argument has less to do with the issue of whether or not we misrepresent, exaggerate and distort the capacities of real human agents in presenting them as (or recommending that they be) HHAs, and more to do with a misplaced emphasis on the more logical aspects of our mentality, and our use of language, at the expense of more imaginative, creative, and capacity- or resource-transcending dimensions of our socially-embedded, symbolically-endowed minds. Although these claims are, strictly, independent of any claim that we might be something significantly less than HHAs, they do fit naturally into a sketch of human agency that intends to avoid distorted images of human agents in which explicit articulated reflection and logical reasoning predominate over less linguistic (e.g. body-centred), more creative (e.g. imaginative), and (where appropriate) extra-cranial (e.g. external memory banks, distributed cognitive networks) dimensions of mind and agency. This might sound somewhat telegraphic at the moment, but that is unavoidable – giving content to these last-mentioned ideas must await the discussions and arguments of Chapter 10. For present purposes, it is sufficient to note that the drive to avoid the image of an HHA is not solely based on evidence and arguments about what we are not and ought not to be *qua* agents; it is also driven by concerns that, in characterising ourselves as something (too much) like HHAs, we may have neglected to highlight some features of human agents that represent real and significant sources for degrees of freedom in action.

That, then, is my basic argument for reshaping the framework in which we understand both the puzzles and the positive projects associated with free agency. We are faced with the challenge of situating free agents somewhere in the space between agent automatons, who we think are not good candidates for qualifying as free agents, and hyper-rational, hyper-

reflective agents, who may or may not be free agents, but we think that we at least are not such agents nonetheless. Our subject matter is the same – we are still trying to pin down the details of, and then defend, our claim that we are free agents – but the subject matter is no longer to be fundamentally structured and constrained by the traditional concerns and tensions surrounding determinism, compatibilism, incompatibilism, libertarianism, and hard incompatibilism/ determinism. We are, as per the title of the current chapter, trying to change the subject without changing the subject, trying to inject movement and fresh directions for the investigation and defence of human free agency into a well-trodden problem space that has come to be dogged by stalemate and impasse, without pretending thereby to have settled the debate or to have produced a definitive account of free agency.

With this alternative framework in mind, it is now time to add further flesh to the figure of the AA by examining the empirical case/s for greater scepticism about the role, relevance and reach of consciousness and conscious will in human behaviour – the subject matter of Chapter 6.

Chapter 6

Consciousness, Automaticity and Illusion: Are we just Agent Automatons?

Over the last fifteen years or so, while the philosophical and interdisciplinary study of consciousness has experienced a notable revival in both interest and ink, the role and reach of conscious agency has come under fire from a number of directions. A number of researchers (especially from outside of philosophy) have offered various pieces of empirical evidence, interpretations of empirical findings, and empirically-inspired hypotheses, that have been held up as challenges to what might appear to be some of our most cherished ideas about human agency. Prominent among these are empirically-motivated concerns about consciousness and its role in the choice, initiation and control action. I have selected three examples of these kinds of empirically-motivated challenge to conscious agency, both because they bear directly on some of the issues raised by O'Connor (2000) as to the proper place of considerations about conscious awareness within accounts of free agency, and because they comprise a cluster of, as it were, operationalisations or concretisations of the threat of the agent automaton (AA) that I sketched in Chapter 5 as part of an alternative framework in which to situate both puzzles and positive accounts of free agency.

The three examples are (i) evidence for automaticity in human functioning; (ii) evidence and interpretations of the so-called readiness potential and the timing of conscious volition; and (iii) Daniel Wegner's account of the illusion of conscious will. Each will be presented in a reasonably uncritical fashion so as to capture the flavour and relevant empirical evidence. A sustained critical evaluation and response will become a significant part of the work for Chapter 7 and subsequent chapters.

Automaticity

A common theme in empirically-motivated discussions of free will is that, in contrast to O'Connor's complaint, our preferred views of human agency are built on strong assumptions concerning the nature, degree and extent of conscious volition and control in our lives. Unsurprisingly, psychologists studying automaticity have suggested that evidence of pervasive automatic, non-conscious functioning might, at the very least, considerably restrict

the extent to which our preferred view of agency is relevant to understanding human behaviour. That is, evidence of automaticity is offered as grounds for thinking that we exert far less conscious influence and control over our lives than many might think; and, further, that our concepts of human agency may require revision in the light of such evidence.

For example, Bargh and Chartrand (1999) present the results of a number of studies that they think suggest a far more limited role for conscious initiation and control than is apparently assumed by those outside of academic psychology. Specifically, they offer evidence of so-called automaticity in the activation of behavioural dispositions, goals, and moods and evaluations, that together suggest a much more modest and restricted role for conscious volition than we might like to think is the case.

First, there are issues arising out of the apparent link between perception and thought, and behaviour – the empirically-supported idea that thinking of or perceiving an action can result in an agent themselves being more likely to perform that action. For Bargh and Chartrand (1999), the notion that perception is a largely automatic process outside of conscious control leads them to surmise that the environment of an agent can influence and control specific behavioural tendencies and dispositions of the agent through the perception-action link, thus bypassing processes of conscious volition and control. They cite a number of social psychological laboratory studies to flesh out their claim.

For example, the three studies conducted by Bargh, Chen and Burrows (1996) each involved the idea of non-conscious priming of traits or stereotypes. In the first study, participants were given two tasks, each supposedly unrelated to the other. In the first phase – a language-related task involving scrambled sentences – participants were exposed to words associated with either rudeness, politeness, or with neither (the control condition). The hypothesis guiding the experiment was that the rudeness- and politeness-related words would non-consciously prime behavioural tendencies in the participants consistent with the trait that had been primed. In the second phase of the experiment, participants were placed in a situation where, in order to receive instructions for (what the participants thought was) a second experimental task, they could interrupt a conversation between the researchers. The results were significant and suggestive: 67% of the ‘rudeness’ group interrupted the conversation,

versus 38% from the control condition and just 16% from the 'politeness' group.[135] For Bargh and his co-workers, the results suggest that mere exposure to words connoting a particular trait can make behaviour in accordance, or consistent, with that trait more likely.

In their second study, Bargh *et al*. (1996) hypothesized that activating a specific stereotype might incline participants to behave in accordance with features of that stereotype. In an apparent language test involving scrambled sentences (phase one of the study), participants were either exposed to words relating to the stereotype of the elderly (the experimental condition) or words with no specific connections to such stereotypes. In the second phase of the study, the behaviour of the experimental participants was unobtrusively observed as they walked down the corridor, and compared to that of the controls. It was found that the participants whose stereotype of the elderly had been activated behaved (i.e walked) in a manner typical of the stereotype. On average, the experimental group took about 1 second longer to walk down the corridor compared to the controls,[136] and were more forgetful of details about the room in which the first phase of the study had been carried out.

In their third study, Bargh *et al*. (1996) set out to test the effects of priming a stereotype of young male African Americans, which previous research had associated with hostility, by subliminally presentating a picture of either an African American man or a Caucasian man before participants completed a computer-based task. The latter task had been specially designed and piloted to confirm that it was experienced as boring and tedious. After more than a hundred trials of this task, the computer flashed an error message, followed by an instruction indicating that the participant would need to begin the task again. At this point, the experimenter's assistance had to be sought, and the experimenter initially confirmed that the participant would need to start over again, before finally indicating that this would not be necessary. Hostility ratings were obtained from the experimenter, based on this interaction, as well as from video recordings of the participants facial expressions[137]. Bargh *et al*.'s (1996) analyses indicated significantly greater hostility in the responses and behaviour of

[135] Bargh *et al*. (1996) originally measured the effect using time as the dependent variable, with participants given up to 10 minutes to interrupt the researchers. But while their ANOVA and post-hoc comparisons supported their hypothesis, the data was not really suited to the analysis since almost two thirds of the sample did not interrupt at all – meaning they had the same score with no variance. Bargh *et al*. (1996) thus followed up this time-based analysis with a comparison of proportions who interrupted under each condition, as reported here, which when tested confirmed a significant linear trend from politeness through no priming to rudeness.
[136] T-tests showed these differences to be statistically significant.
[137] All hostility ratings were blind to the priming condition of the participant being rated.

participants in the male African American stereotype condition, suggesting that subliminal priming of this stereotype was sufficient to activate behavioural tendencies consistent with the stereotype[138].

In a related study by Chen and Bargh (1997) where the same priming technique and stereotype were used, participants primed for the male African American stereotype (experimental condition) were judged[139] by independent raters to have shown greater hostility towards their partners in a word-guessing game designed to be mildly frustrating[140], as compared to those primed with a picture of a Caucasian male (control condition). Participants in the experimental condition also managed to provoke greater hostility from their fellow participant during the guessing game.

The second domain of automaticity focussed on by Bargh and Chartrand (1999) involves goals and goal activation. Their basic claim is that the regular activation of a particular goal in a given situation can, over time, lead to the automatic (hence non-conscious) activation of that goal by salient features of the situation. Bargh and Chartrand cite evidence from their own studies where cognitive goals were claimed to be activated non-consciously by way of priming, instead of via explicit instructions, while nevertheless yielding the same consequences for performance as one would have expected from explicit goal activation. For example, participants in a study were given a list describing various behaviours of an agent, but were not given explicit instructions as to what they should do with this information. Prior to this activity, however, each participant had been engaged in a language-based task during which they had either been exposed to words associated with memory and memorising (such as 'retain' or 'hold'), or to words connoting evaluation (such as 'judge' and 'evaluate'). Participants in the study produced the same overall pattern of recall and memory organisation

[138] Bargh *et al.* (1996) also administered two racism scales to their participants to test for any possible covariation between the observed hostility effects and pre-existing racist attitudes. No significant correlation was found, suggesting to these authors that the association between the stereotype and hostility, as well as the effects observed in the study, were not significantly associated with or moderated by consciously expressed racist attitudes.

[139] Again, rates blind to the priming condition were used – this time based on audio recordings of the interactions during the word-guessing game.

[140] The game was mildly frustrating in the way that many guessing games are – the participant giving clues for a given word was restricted in the kinds of clues they could use. Judges ratings of hostility under these circumstances were thus tailored to the kind of frustration likely to emerge in such a game: for example, a hostility rating of 5 (out of 7) was described as "Significant signs of frustration. Characterized by outward annoyance, but still attempts to remain civil.", 6 as "Display of moderate outward hostility. Heightened voice level, significant outward annoyance, signs of anger.", and 7 as "High levels of outward hostility. Yelling, use of insults, and derogatory comments." (Chen & Bargh, 1997, p552).

for these different cognitive goals as had been obtained in earlier studies that had used explicit instructions to either memorise the behaviour information or form an impression of the agent.

Bargh and Chartrand (1999) provide further illustrations of cognitive goal activation by citing a group of studies by Spencer, Fein, Wolfe, Fong and Dunn (1998) that were based on the idea that threatening someone's self-image tends to automatically prompt attempts at restoring their self image by a variety of means, including the denigration of others. One means towards cognitively denigrating others is by employing negative or derogatory stereotypes. In their study, Spencer *et al*. (1998) claimed to have shown that techniques known to reduce or eliminate the use of stereotypes were rendered ineffective – that is, the stereotypes remained active – in participants whose self-image had been threatened by way of negative feedback on task performance.[141] Bargh and Chartrand (1999, p470) interpret the significance of this study as demonstrating that "…the threat to self-esteem put into motion a goal to denigrate others that was so automatic and efficient in its working that it produced stereotyping of a minority group member under attention-overload conditions, in which manifestations of stereotyping are normally not obtained."

Even behavioural goals are susceptible to automatic (situational) activation that would appear to bypass processes of conscious initiation, oversight and control. Bargh and Chartrand (1999) overview the findings of a number of studies in which the possibilities for priming goals such as achievement motivation were examined. For example, participants in an experimental group performed a word search task in which they encountered terms associated with achievement (e.g. 'strive', 'succeed'), while controls performed a similar task that did not include achievement-orientated words. In subsequent tasks, the experimental group outperformed the controls, without evidencing any awareness of an association between or effect of the earlier priming task on subsequent tasks. Similarly, in another study in which experimental participants had their achievement goal primed, these participants continued

[141] So, for example, in Spencer *et al*. (1998) Experiment 1, participants performed a word completion task under cognitive load (previous studies having suggested cognitive load reduced or prevented the activation of stereotypes). Control participants showed no evidence of stereotype activation, whereas experimental participants who had received negative feedback just prior to the word-completion task showed evidence of stereotype activation in the words they offered despite performing under cognitive load. In this study, the act of denigrating others thus amounted to offering words during the task that were associated with stereotypes of Asian Americans under the experimental condition in which an Asian American woman was shown in a video holding up the stimulus cards for the word-completion task.

with a word generation task[142] beyond the allowed time limit to a significantly greater extent than did non-primed participants (55% versus 21%).

According to Bargh and Chartrand (1999), the effects of this non-conscious priming of behavioural goals extend beyond just task performance to include the consequences of success or failure for things like mood and self-efficacy. For instance, they report on Chartrand's unpublished findings in her doctoral research in which she, once again, primed experimental participants' achievement goal, and then examined the effects of manipulating task difficulty in an anagram task for which participants were told they were given an average amount of time. Measures of mood after task completion showed that control participants were unaffected by variations in task difficulty, whereas experimentally primed participants were in a significantly worse mood after the difficult anagram task compared to the simple one.

As in all of this research, the evidence further suggested that none of the participants reported any specific achievement goal on the experimental task. For Bargh and Chartrand (1999), this shows that people can be moved, non-consciously and automatically, to pursue a cognitive and/or behavioural goal, and respond appropriately at a psychological level to the attainment or frustration of that goal, without awareness that these processes are in motion.

Rather more unpleasant evidence of situational goal activation is offered in a study by Bargh, Raymond, Pryor and Strack (1995) on males identified as being prone to sexual harassing or sexual aggression. In a priming task measuring the time to pronounce (out loud) a word presented on a screen following the subliminal presentation of a prime word (the priming words being related to power, sex or neither), males identified as likely sexual harassers or aggressors were the only participants to respond more quickly in pronouncing a sex-related target word following a power-related prime word. No corresponding decrease in response time was observed when a power-related target word followed a sex-related prime. Bargh and Chartrand (1999) interpret these results as suggesting that, in these individuals, a situation involving power automatically activates concepts of sex. While it strikes me that there is something of a leap to be made from activation of *ideas* about sex to activation of sexual

[142] The task required participants to write down as many words as possible from seven Scrabble letters in a three-minute period. At the end of the three minutes, a clear 'Stop' instruction was given by the experimenter, and hidden video cameras were used to record any attempts to continue with the task after this instruction had been received.

goals, the trend of the findings nevertheless lends further credibility to the authors' claims about the potential power of situational factors to non-consciously connect with an individual's conceptual and motivational systems.

With regard to their third area of focus – evaluations – Bargh and Chartrand (1999) again offer evidence that the processes involved in generating these are much more automatic than some might like to think. In contrast to a reasonably widespread view of emotions and moods as occurring with little conscious choice being involved[143], Bargh and Chartrand (1999) argue that evaluations are generally thought of as being made consciously and intentionally:

> Many theories of attitude formation and of the evaluative process hold that one weighs the pros and cons, or positive and negative features of the object or event, and with intention and deliberation makes a decision about how one feels about it… (Bargh & Chartrand, 1999, p473)

In contrast to such views of evaluation, Bargh and Chartrand (1999) outline a tradition of theory and research in psychology that suggests a view of evaluation, and of the activation of evaluations, as being largely automatic and outside the realm of conscious intention. Within this tradition, people are viewed as showing preference and affective-evaluative responses automatically, and specifically in advance of any awareness of reasons for such preferences and evaluations.[144]

Examples of empirical findings supporting a more automatic view of evaluation (and its impact on mood, behaviour and conscious judgments) cited by Bargh and Chartrand (1999) are too numerous to describe in detail. They include (i) priming tasks where automatic evaluations of primed concepts are said to explain faster response times to positive or negative adjectives; (ii) the apparent influence of performing verbal tasks involving nouns with widely-shared positive or negative evaluative associations on subsequent measures of mood; (iii) interactions between automatic evaluations and behavioural dispositions such that, for example, individuals would be slower to respond to a positive stimulus when the required response was the pushing of a lever (a generally avoidant bodily movement) than when the required response was the pull of a lever towards the body (an 'approach' response); (iv) evidence that predictions of the character and behaviour of strangers was at least as accurate when based on 'immediate' responses (observations of 30 seconds or less) as when based on more extensive observations and conscious deliberation; and (v) evidence that the longer an

[143] Claims about the speed, automaticity, alleged 'passivity' and relatively primitive status of emotional response are critically discussed by, amongst others, Damasio (1994) and Blackburn (1998).
[144] These accounts of links between perception and affect clearly resonate with Damasio's (1994) claims about somatic marking.

individual consciously considers and deliberates over their evaluations and judgements, the more likely it is that accuracy and predictive value will be *lost*.

To what, then, does the "unbearable automaticity of being" of Bargh and Chartrand's (1999) title refer? By their own lights, the central thesis of their paper is "that most of a person's everyday life is determined not by their conscious intentions and deliberate choices but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance" (Bargh & Chartrand, 1999, p.462). This thesis, together with the cited evidence of non-conscious influence, might suggest a number of potentially 'unbearable' conclusions to not only lay people but also various theorists in the debates over free agency. First, what we might label 'regularist' accounts of free agency (Searle's 'gappy' account of choice and action; agent-causal theories like O'Connor's; compatibilists who emphasise action in accord with conscious decisive better judgements or conscious reflective endorsement) should feel the squeeze being exerted on the availability of candidate choices and actions, engaged in with conscious intention and deliberation, that could serve as clear evidence for freedom being a general and regular characteristic of our lives as human agents. If Bargh and Chartrand (1999) are correct in their arguments and their evidence, candidate choices and actions might be 'unbearably' thin on the ground.

Second, many might worry that the evidence for automaticity threatens to contaminate even the remaining candidates for conscious volition and control. That is, once it is allowed that non-conscious influence *can* operate in the ways described, one may worry that the activity of such automatic processes cannot be sufficiently discounted in cases of apparent conscious volition so as to persuade us that these might meet the criteria of conscious intention and control set by the regularist accounts of free agency. Such a concern would have both an epistemic aspect (contamination of possible evidence) and an ontological aspect (e.g. extensive contamination of motivation through parallel activity of conscious and non-conscious processes).

And a third source of concern might be the extent to which the evidence for automaticity could be used to underpin claims of illusion in conscious volition that will be further explored in the context of Daniel Wegner's work (below). As Bargh and Chartrand (1999) note, we are almost by definition unaware of the operation of non-conscious processes. This makes it relatively unsurprising that people might be generally inclined to excessively discount, or

flatly reject claims about the extent of such influence. But if the evidence is to be believed, these objections and denials might be treated as further evidence that consciousness is, in some sense, rich in a variety of illusions, especially when it comes to our consciousness of our own agency.

The Timing of Conscious Volition

For all the concerns that the evidence about automaticity might raise about conscious volition and control, it is at least allowed that consciousness can sometimes be an initiator of behaviour, an activator of goals, and a source of evaluative judgements. Data from studies of neural activity preceding voluntary movements, and innovative attempts to map conscious volitional processes onto the timescale of this neural activity, have been interpreted by some as potentially undercutting *any* claims of conscious initiation of action.

As early as the 1960's, Kornhuber and Deecke (1965, cited in Dennett, 2003 and Libet, 1999) identified a pattern of brain activity, preceding voluntary bodily movements by between 800msec and 1 second, that has become known as the *readiness potential* (RP). The RP (readiness potential) has been widely interpreted as an indication of neural preparation in the motor and premotor cortices for an ensuing voluntary movement, and its discovery has suggested an apparently straightforward question about conscious voluntary behaviour and its relationship to underlying neural activity – namely, how does conscious volition map onto the apparent timeline of neural activity indicated by the RP?

Adapting the methods used in the original studies, Benjamin Libet and his colleagues began investigating this question by asking participants to perform a flick or flexing of the wrist in an unplanned way, whenever they felt the wish to do so (Libet, 1999), and found an average RP of -550msec before the movement. Libet summarises the logic of his studies as follows:

> The brain was evidently beginning the volitional process in this voluntary act [of wrist flicking] well before the activation of the muscle that produced the movement. My question then became: *when* does the *conscious* wish or intention (to perform the act) appear? In the traditional view of conscious will and free will, one would expect conscious will to appear before, or at the onset, of the RP, and thus command the brain to perform the intended act. But an appearance of conscious will 550msec, or more before the act seemed intuitively unlikely. It was clearly important to establish the time of the conscious will relative to the onset of the brain process (RP); if conscious will were to *follow* the onset of RP, that would have a fundamental impact on how we could view free will. (Libet, 1999, p.49; italics in original)

In order to address the critical issue of timing, Libet embellished the laboratory procedure by asking his participants, as before, to flex their wrists as and when they felt the wish to do so,

but all the while observing a moving dot rotating on an oscilloscope at a rate of 2.56 seconds per revolution. Participants were requested to note the position of the rotating dot on the oscilloscope face at the precise moment when they first became aware of a conscious wish to flex their wrist – the position of the dot to be reported after the wrist flexing trial so as not to interfere with the trial itself.

Allowing for an estimated error of -50msec in timing using the above method, Libet found that the average reported conscious wish was -200 to -150msec before the activation of the muscle involved in the wrist flick, and thus 350 to 400msec *after* the onset of RP at an average of -550msec before muscle activation. For Libet (1999, p.51), this indicated that, "clearly, the brain process (RP) to prepare for this voluntary act began about 400msec before the appearance of the conscious will to act…". For Libet and many other commentators, the implications of these results were clear: "The initiation of the freely voluntary act appears to begin in the brain unconsciously well before the person consciously knows he (sic) wants to act!" (Libet, 1999, p.51) Here, it seemed, was an undeniable challenge to any view of the conscious human initiator and 'unmoved mover' behind their consciously willed voluntary actions. Conscious will seemed to make its appearance too late to sustain such a view of conscious volition.

Various aspects of Libet's studies will be discussed at greater length in Chapter 7. For the moment, it is worth noting that despite the apparent simplicity of the studies and their results, it is probably fair to say that there is no widely shared interpretation of the significance of Libet's data. Libet's own interpretation of the results is seldom taken at face value, and his proposal of a conscious 'veto' (between the appearance of a conscious wish and the activation of movement) as the best means to understanding (if not saving) free will is widely rejected. What these empirical data clearly offer, however, is one part of an explanation or motivation for the apparent absence of detailed commitments on consciousness and conscious volition in many accounts of free will – namely that pronouncements on these matters may carry with them the threat (or promise, depending on one's Popperian sentiments) of empirical refutation. Alternatively, the possibility of empirical hypothesis testing of this variety might demand a much more detailed and empirically informed theory of psychological and brain function to accompany an account of free agency than philosophers have seen fit (or necessary) to provide, if only to be better able to explain away any apparent empirical failures that might emerge in the laboratory.

To the uncritical eye, Libet offers us an innocuous choice situation, apparently free of the deceptions, manipulations and proddings that permeate the empirical studies of automaticity reviewed above, leaving the conscious will free to show us its true form. If we share Libet's intuition that his results do not fit with an ordinary notion of the conscious will as an initiator of action, just how much else about conscious will might we have gotten wrong?

Illusions of Conscious Will?

The theory of conscious will offered by Daniel Wegner (2002; Wegner & Wheatley, 1999) suggests that we might, in an important sense, have it *all* wrong because our experience of conscious will is illusory. Of course, it is important to get clear on just what this illusion is supposed to consist in – Wegner (2002) is evidently not a skeptic about mental causation. Nevertheless, his work (and the provocative title of his book *The Illusion of Conscious Will*) clearly demands attention as an important recent example of empirical challenges to our understanding of volition that work, in part, by challenging assumptions about the conscious aspects of agency.

Wegner's (2002) work is strongly influenced by Dennett's ideas about the so-called intentional stance, as well as by empirical and philosophical work on the development of theory of mind. The essence of his theory is that the conscious will – what Wegner (2002) calls the phenomenal will – is an aspect of mind engaged in a process of self-interpretation and explanation of action, whereas what he calls the empirical will refers to the aspects of mind that are actually (causally) related to action – all the causal relations that are the proper subject matter of psychology. The intended contrast between these two senses of will is neatly captured in the following extended quotation:

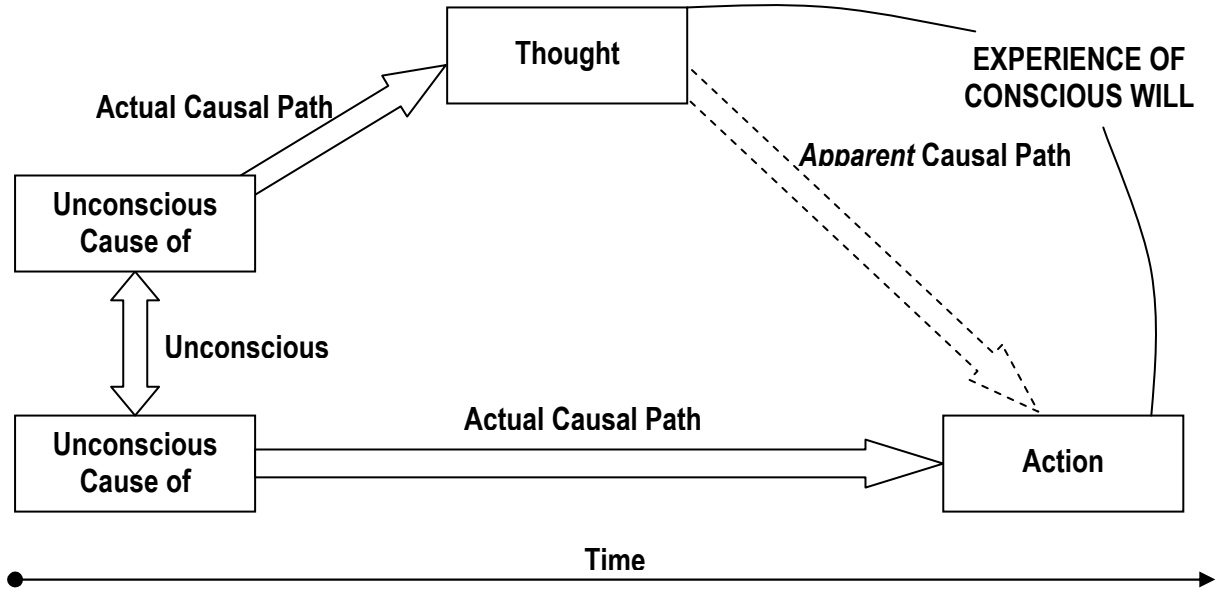
> Whatever empirical will there is rumbling along in the engine room – an actual relation between thought and action – might in fact be totally inscrutable to the driver of the machine (the mind). The mind has a self-explanation mechanism that produces a roughly continuous sense that what is in consciousness is the cause of action – the phenomenal will – whereas in fact the mind can't ever know itself well enough to be able to say what the causes of its actions are. (Wegner, 2002, p.28)

At one level, these ideas fit neatly into a tradition in philosophy that includes Spinoza and Schopenhauer (Wegner (2002) himself notes a connection to Spinoza) in which it has been argued that consciousness remains ignorant of many if not most of the actual causal factors underpinning action, and yet such ignorance is then mistaken for freedom from (Spinoza) or

the absence of (Schopenhauer) such causal influences.[145] Wegner's (2002) take on these matters involves offering a general theory of how the experience of conscious will arises in which the crucial claim is that the phenomenal will lies on a causal pathway that is ultimately disconnected from action, such that the experience of conscious thoughts having caused an action is always illusory (or, to use the more neutral terms Wegner (2002, p.2) offers in a footnote, a construction or fabrication).

The overall structure of the theory is perhaps best represented in the following figure, adapted from Wegner and Wheatley (1999, p.483):

Figure 6.1: Wegner's Model of Conscious Will (adapted from Wegner, 2002)



Wegner (2002; Wegner & Wheatley, 1999) subscribes to a Humean view of the conditions under which we infer causal relations, and the experience of conscious will is then treated as a special case of causal inference establishing an apparent causal connection between a conscious thought (most often a conscious intention preceding action) and an ensuing action. More specifically, given (i) an appropriate degree of temporal precedence, (ii) an appropriate degree of consistency between the thought (content) and the action, and (iii) an appropriate degree of exclusivity of the thought as the only (or most obvious) candidate for being the cause of the action, the mind will experience the conjunction of the conscious thought and

[145] Links to Spinoza and Schopenhauer are also to be found in Blackburn's (1998, 1999) work.

ensuing action as a causal relationship. As is suggested in Figure 6.1., the causal origins of both the conscious thought and of the action are unconscious. Two parallel causal pathways diverge from this 'point' of unconscious initiation of thought and action, one leading to the conscious thought and the other leading to the performance of the action, but crucially never intersecting or converging. The illusion of conscious will thus arises from the notion that the apparent (inferred and experienced) causal pathway from conscious thought to action is *never* the actual causal pathway that culminates in action.

I take it that this is the crucial claim in Wegner's theory, and provides his most significant motivation for choosing the provocative and polemical term 'illusion' as his preferred label for the phenomenon. Conscious will is not illusory because we do not have such experiences, or because we are sometimes mistaken about the content of such experiences. Wegner wants to argue that our experience of conscious will is, on the whole, an important and useful aspect of how we function as agents in the world. Moreover, he seems to allow that when the unconscious causes of action and conscious thought are suitably related, the conscious thought (especially conscious intentions) will accurately flag the action that follows, and may also provide a (partial and incomplete) representation of the causal origins of the action in the psychology of the agent. What remains illusory under even these ideal circumstances is the inference (and experience) of the conscious thought being a part of the *actual* causal pathway leading to action because, given Wegner's (2002; Wegner & Wheatley, 1999) model, this inference is never correct.

Wegner draws on a wide range of empirical data as potential evidence for his theory. For instance, he thinks that Libet's findings fit very neatly with his model of the conscious will (see Wegner & Wheatley, 1999, and Wegner, 2002). However, the majority of the evidence he gathers (especially in Wegner, 2002) can be understood as an attempt to establish, in something like the tradition of neuropsychology, a pattern of dissociations between the experience of will and the origins of action that is consistent with his theory. On one hand, he needs to provide evidence of demonstrably intentional action that is not experienced as having been intentionally caused. On the other hand, he needs evidence that the experience of consciously intending or willing an action can be generated under circumstances where the candidate 'action' was demonstrably not intended by the agent. With these dissociations in place, Wegner will offer his parallel pathways model as the most plausible explanation for the relevant phenomena – that is, you can explain the dissociation of experienced

intentionality and actual intentional action best if you assume that these phenomena are always distinct.

Wegner's (2002) preferred evidence for intentional action in the absence of experienced intention comes from the study of what he calls automatisms. While he includes some of the phenomena discussed by Bargh and Chartrand (1999) under this banner, Wegner's (2002) favourite examples are of a more exotic variety – including Ouija boards, spiritualist table-turning, and automatic writing. Consider, for example, the nineteenth century phenomenon of table-turning reported at various spiritualist séances. Participants in these séances claimed that the table was being moved by the activities of spirits trying to communicate with the gathered (chosen?) few (Wegner, 2002). Not only was there no apparent awareness or consciousness of the participants being responsible for the movement of the table, but those attending expressed the conviction that they were *not* the source of the movement. Yet, Wegner (2002) describes how Michael Faraday was able to demonstrate, through the use of devices attached to the hands and the table to measure force, that the origins of the table movements clearly lay in the collective actions of the participants (Wegner, 2002, claims this can be established equally well be examining the direction of finger streaks left on a dusty table). To Wegner, this suggests that we can find instances of demonstrably intentional action (action arising out of the desires and expectations of the séance attendees), actively produced by agents, and yet produced without being accompanied by the experience of intentionality and voluntariness.

In order to establish the other dissociation implied by his theory, Wegner (2002) harnesses the three conditions offered for the experience of conscious will to predict that a feeling of having consciously intended an action could be manufactured – an illusion of conscious control could be generated – in cases where it was clear that the agent's intentions were not the source of the action performed. As a mundane example of an illusion of control, Wegner (2002) describes the experience of apparently controlling the movements of characters in a computer or arcade game, only to have "Game Over" or some instructions as to how to start the game flash on the screen, thus revealing the illusion. Through a combination of timing of movements, sufficient coherence between intended movements and the movements of the game character, and the relative exclusivity of your intentions and movements of the controls as an explanation of the 'game's' unfolding, you may have a strong conviction that the

character is responding to your actions – you are in control. When it becomes evident that the game has not started, the illusion is revealed in an instant.

A more refined example of the illusion Wegner has in mind comes from laboratory studies of the experience (and malleability) of intentional control. Wegner and Wheatley (1999) describe a study designed to investigate whether an illusory experience of intention could be generated for behaviour that was in fact performed by another person. Two participants in the study (one was a confederate) were seated in front of a computer monitor, their fingertips placed on a Ouija board-like device (a board on top of a computer mouse). On the screen was a collage-like picture of approximately 50 small objects (e.g. car, swan, dinosaur) and a cursor whose movements were controlled by movements of the board/mouse. The apparent procedure was for the two participants to move the cursor around the screen for about 30 seconds, at which point they would need to make a stop. Each stop would then be rated for the degree of intentionality felt by the participant (on a scale from 'I allowed the stop to happen' to 'I intended to make the stop'). Both participant and confederate wore headphones, and the participant was under the impression that both would be hearing a mostly synchronised but different soundtrack – namely, 30 seconds of silence, followed by 10 seconds of music indicating that a stop should be made. In addition, words would occasionally be heard over the headphones, and these were explained as providing a distraction, with the participant and confederate hearing different words.

In reality, the confederate heard neither music nor words, but instead received instructions (on the key trials) to move to a particular object on the screen, followed by a countdown (overlapping with the music being played to the participant) to them attempting to force a stop of the cursor on the named object. On these key trials, the participant heard the word for the object given in the confederate's instructions at different times relative to the forced stop: 30 seconds before, 5 seconds before, 1 second before, or 1 second after the stop.

For all other trials, the confederate allowed the participant to make the stops as they wished. In about half of these trials, the participant heard a word for an object on the screen, whereas for the other trials the word had no corresponding object. Hearing a word naming an object on screen was intended to act as a candidate conscious thought that could satisfy the three conditions for being experienced as the cause of an action. Comparisons of the unforced trials suggested that hearing the word for an object on the screen did not lead to stops that were

significantly closer to the named object than when there was no word-object correspondence. Since hearing a word for an object did not, on its own, lead to significantly more stops on or near the object, the researchers inferred that the forced stop trials were unlikely to have involved collusion on the part of the participant.

In the forced stop trials, participants tended to report the stops as intentional, but the effect varied according to the timing of the 'prime' word for the object on the screen. When the word was heard either 30 seconds before or 1 second after the forced stop (thus undermining the temporal condition of experiencing conscious will), participants on average reported less intentionality; when the word was heard 5 or 1 second before the stop, much higher ratings of intentionality were reported (specifically, ratings in the same range as those given for the unforced trials where the stop had been determined by the participant).

For Wegner (2002; Wegner & Wheatley, 1999), this study provides evidence under controlled conditions for an illusion of control that provides the other dissociation predicted by his model of conscious will – the experience of intentional, conscious willing under circumstances where (through the contrast with relevantly similar cases) it seems the real explanation lay in the behaviour and intentions of the confederate. More significantly, perhaps, this case suggests evidence for the dissociation that is more difficult to establish. It is probably the case that people are, in general, not averse to discovering and owning various intentional aspects of action after the fact of which they were unaware at the time of performing the action – acting intentionally without awareness of intention. The permeation of psychoanalytic thinking into popular or folk psychology arguably provides evidence of a willingness to accept retrospective intentional explanations of action that were not obvious to (indeed, might have been denied by) the agent at the time of action. Providing convincing evidence of an illusory sense of intentionality and conscious control seems to be a much more demanding task, making the generalisability of Wegner's laboratory evidence a critical issue in the evaluation of his theory.

Taken together, Wegner has offered evidence for dissociations between the experience of will and what he calls the empirical will that attempt to drive a wedge between our experience of volition and the underlying reality of mental causation. As indicated above, the alleged illusion of conscious will is not an illusion because we are necessarily out of touch with our intentions and other mental causes of our actions. When optimal conditions obtain, our

experience of conscious will may provide us a partial and useful insight into the causal origins of our acts. The illusion Wegner describes is supposed to lie in the fact that the causal chain we experience as the source of our actions is not merely incomplete, but rather is *never* the causal chain at work in producing behaviour.

Chapter 7

Mental States, Processes, and Conscious Intent

What should we make of the various empirically-based challenges to conscious will presented in the previous chapter? Do they provide further motivation for avoiding commitments about the role of consciousness in free agency? Or should we share O'Connor's (2000) view that any adequate account of free agency must address issues relating to the role of consciousness, such that the empirical evidence cited above only raises the odds stacked against our succeeding in such a task?

As suggested in Chapter 5, my own view is that O'Connor is correct to insist on the importance of consciousness, in part because an adequate account of free agency should be able to successfully rebut any objections to the claim that we are free agents arising out of precisely the kinds of empirical evidence reviewed in Chapter 6. If O'Connor (2000) is correct in his assertion that available accounts typically neglect to give or find a significant role to/for consciousness, it is not immediately apparent how these challenges could be rebutted except by way of some unsatisfactory side-stepping of the challenges involved. Taken together, the evidence offered for the pervasive role of automatisms, for readiness potentials that 'jump the gun' on conscious intentions, and for the claim that our experience of conscious will is always illusory must surely undermine any confidence we could have in our experience of agency to provide relevant and credible data on which to base a theory of free agency, and data to which any such theory must answer.

In short, the empirical studies of conscious volition described earlier suggest that conscious volition is not what it would appear to be from a phenomenological perspective. If this is right, then it becomes unclear what the relevant explananda are for an account of free will that takes seriously the challenge of addressing itself to occurent mental phenomena. It also becomes unclear what sources of intuitions about free agency we can rely on to test the implications of available accounts. When we put consciousness and conscious volition under the microscope of laboratory study, it would seem that consciousness constantly moves out of the frame, or gets sidelined, as soon as we push for details about the origins, timing and precursors of choice and action. Of course, it may turn out that conscious volition is not what

we thought it was, without it being the case that consciousness can be sidelined (or rendered prone to ubiquitous illusions) in the ways suggested by some interpretations of the empirical evidence we have considered. One way or another, there is a strong *prima facie* case for moving consciousness and conscious volition more centre stage in the debate about free agency.

My primary concern, for the moment, will be with examining the sense in which automaticity, the timing of readiness potentials, and the apparent illusions of conscious will, might individually and jointly comprise a threat that can be understood under the rubric of AA – the spectre of our being Agent Automatons.

Two strong themes run through the three areas of research we have surveyed. First, each set of ideas (and related empirical evidence) in some way presents a domain of apparently voluntary activity where, so it is claimed, we *think prima facie* that some kind of conscious mental activity or process lies at the beginning of the causal chain that ultimately leads to action, when in fact the evidence suggests that there either is no such conscious event or process, or if there is, there is evidence and reason to think that it does not or cannot play the initiating role we would have it play. This is the core threat of AA – that while we think we have a certain kind of role consciously directing and steering our course through our lives, there is mounting evidence to suggest that we don't and/or can't play that role. Instead, we are more like some kind of automaton running along on (unconscious) autopilot.

Second, and mostly because we have this mistaken sense of having conscious control where we are, it seems, actually running along as if on unconscious autopilot, our conscious self-understanding is in various ways flawed, illusory and misguided. Isolating and highlighting this theme helps raise serious questions about the extent to which we can use our apparent self-knowledge, introspective reports, and experience-based intuitions, as suitable data with which to test out claims about freedom and agency. We are not only AAs, but AAs who don't know it – indeed, who refuse to acknowledge it.

As suggested in Chapter 5, it is possible to respond to the threat of AA, and/or to the specific claims and evidence presented in Chapter 6, by proposing that dealing with such threats is not a priority within the context of the traditional free will debate. From such a perspective, claims about automaticity, or the timing of conscious intentions, or supposed illusions of

conscious will, do not require urgent attention within the context of a traditional debate about free agency in so far as these puzzles (if they are indeed puzzles) need to be addressed and dealt with by *both* compatibilist and libertarian alike. Until it can be shown that this material presents some *differential* degree of challenge for the opposing sides of the traditional debate, addressing these issues might not be a priority because it will not necessarily advance the cause of one side or the other in their debate on free agency.

The reasons for rejecting this kind of response should, by now, also be familiar. That any particular issue might not generate movement within a debate that tends towards an unproductive impasse does not provide a good reason for thinking that the issue is not pressing or urgent. Those 'reflective agnostics' with an interest in defending our claims to being free agents are likely to be troubled by the claims discussed in the previous chapter, and will want a response to especially the more sceptical and threatening of these claims irrespective of whether or not that response also advances the cause of either libertarianism or compatibilism.

Resisting the Threat: Events and Processes, Timing and Time-scales

There are many individual issues, claims and interpretations of evidence at stake in the work that has been discussed. In what remains of this chapter, I will only begin to offer the first of a series of responses that will unfold in the latter parts of this thesis. My initial focus will be on Libet's work, partly because it is the more familiar and widely discussed data within philosophical discussions of agency[146], and because some of what needs to be said about Libet's work will have relevance to the automaticity literature and (especially) to Wegner's account of conscious will.

As we have seen, Libet's work was apparently motivated by a very simple idea: if we can record and time the onset of the brain activity that precedes an instance of voluntary behaviour, and if we can time the onset of the accompanying conscious mentation, we can relate the timing of the one (neural activity) to the other (conscious intending). The first thing to be said about Libet's (and other similar) studies is, perhaps unsurprisingly, that this apparent simplicity is misleading. There is, needless to say, an extensive critical literature

[146] There is a vast and ever-expanding literature on Libet's work, including extensive discussions in the pages of *Behavioral and Brain Sciences*, and in the *Journal of Consciousness Studies* special issue of 1999 titled *The Volitional Brain*. I cannot hope, nor do I intend, to survey or discuss this work in any amount of detail, as it would only serve to distract us from the central points of clarification and responses that I wish to make.

built around Libet's work, and I cannot and will not pretend to offer an extensive review of that work here[147]. But some important points of clarification must be made about the very basic ideas Libet thought were being put to work in his experiments.

The initial points of clarification I have in mind follow from two general correctives about how we think and talk about the mind, mental states and processes. The first of these is that mental states are not, in general, *things* that we *have*, or discrete entities, events or states that we are somehow *related to*. Instead, they are states that we *occupy*.

Galen Strawson (2004, p288) alludes to the need for such a corrective when he laments "philosophers… [talking] in a strongly reificatory way about mental states as if they were things in us, rather than things – states – we are in." And Simon Blackburn hints at the need for a similar corrective when he asserts that:

> Typically, in deliberation what I do pay attention to are the relevant features of the external world: the cost of the alternatives, the quality of the food, the durability of the cloth, the fact that I made a promise. I don't also pay attention to my own desires… Deliberation is an active engagement with the world, not a process of introspecting our own consciousness of it… [Consider] the many ways of failing that await the poet who makes his or her own consciousness of emotion into the subject of a poem, instead of the emotion itself. (Blackburn, 1998, p. 254)

Strawson and Blackburn are making somewhat different points. Strawson's (2004) complaint is quite general, and is directed against any tendency to turn mental states, *qua* states that we are in or that we occupy as whole persons or agents, into reified internal mental entities in us that we somehow have and that we, thus, must bear some relation to. So if I am in the mental state of, say, intending to get myself a cup of coffee, it should *not* be inferred that there is necessarily a mental entity of some kind in me – namely, an intention whose content is that I should get myself a cup of coffee – to which I am somehow related. Intention is a state that I am in, that I occupy, as an agent and as a subject of experience. Of course, we talk about having intentions, just like we talk about having beliefs, or desires, or various emotional states. But this talk of 'having' is part shorthand and part grammatical artefact. When we say we 'have' any of these mental states, what we mean (or should mean) in the first place is that we are *in* these states, rather than that 'they' are necessarily anything *in us*.

[147] See, for example, peer commentaries and responses accompanying Libet (1985), as well as various contributions to Libet et al (1999). Dennett discusses Libet's work extensively in Dennett (1991), and revisits some of what he regards as the central issues in Dennett (2003). Libet defends his data, methods and interpretations in Libet (2004).

Blackburn's (1998) claim is being made in the context of criticising certain assumptions about deliberation and the deliberative position[148]. His point is to highlight that our mental state in the midst of a given deliberative process is *world-directed* rather than being somehow introspectively directed at the contents of our minds, or at our consciousness of the world and our consciousness of the contents of our minds. When we deliberate, what we contemplate is the world *as shaped by*, as structured by, our interests in particular (in terms of what we notice and what appears salient), and by our psychologies more generally. For Blackburn, it is a mistake to think of deliberation as some largely or significantly introspective process that involves (requires) second-order reflection on our desires, values, beliefs and intentions.[149]

Whatever the full impact of Blackburn's claims on issues of deliberation *per se*, we can note for the moment that the spirit of his claim is in significant agreement with Strawson's (2004) concerns. If we read Blackburn in the terms offered by Strawson, we could say that to 'have' a given desire or interest[150] (that is at play in our deliberations) is to *occupy*, to *be in*, a mental state that is world-directed; it is not to have a mental state (the desire, the interest) that is necessarily some thing in us, that we introspectively examine, and that gets pushed and pulled around as we deliberate.

The second general corrective I have in mind relates not so much to mental states as it does to intentional behaviour – to actions. If we are prone to talk about mental states as if these were somehow discrete things – mental entities – that are somehow in us, then we are also inclined to talk about the actions we engage in (with intention[151]) as if these were discrete, isolatable and pin-pointable events in the world (including, in the case of 'mental' actions like choices, in us or in our heads). But very few actions can sustain this push towards discrete identification of beginnings, ends and other boundaries.

Consider some of the classic examples that populate the philosophy of action – match strikings and assassinations. Match strikings might look like they have discrete beginnings and ends, but these can be made to look more or less illusory. For a start, any appearance of

[148] We will return to his discussion of these issues in Chapter 9.
[149] This is not to say that we are not capable of introspection and second-order reflection.
[150] 'Interest' is Blackburn's preferred term when talking about deliberation, rather than the more specific term 'desire'. See Chapter 9 below for discussion.
[151] In so far as they are actions, it can be assumed that these behaviours are by definition engaged in with intention. I am simply making the point explicit upfront, given the prominence of notions of intention and intent in Libet's work.

distinct boundaries tends to be a function of timescale, and if we magnify our analysis to an ever finer timescale, we will land up with an endless series of essentially arbitrary choices of beginnings and ends. Moreover, the boundary between the event that is the supposed action of match striking and that which is, say, the preparation for match striking following a decision to strike a match, and the boundary between the supposedly discrete event that is the decision and all that follows, suffer from the same sense of potentially arbitrary discreteness being imposed on things which are inherently extended in time and space, and fuzzy around the edges. The same goes for assassinations and most of the other less exotic, everyday actions that people engage in.

For the most part, then, actions are (like consciousness) *processes* that are inherently extended in time and space. They are not naturally discrete entities or events, nor are they easily or sensibly decomposed into such discrete entities or events, nor into series of discrete events (or micro-events). They exist and evolve at an ontological level that maps both closely enough and yet also loosely enough onto events and processes that we might carve out in various other sub- or non-intentional vocabularies. And when we are not doing philosophy, or engaging in certain kinds of neuroscientific or psychological experiments, this desirable and appropriate looseness of fit does not worry us in any significant way[152].

The significance of these two correctives to Libet's work should be clear. For one thing, the apparent simplicity of Libet's original idea of mapping the timing of conscious intentions onto neural 'preparations for action' should already look suspicious in the light of the above considerations. Conscious intentions are not just 'things' that somehow happen in or arise in us at discrete instants in time. Instead, we enter and occupy states of conscious intending whose boundaries in time are fuzzy, continuous and overlapping with a range of other mental and bodily processes that evolve over time. Exactly where a supposedly discrete 'first' awareness of the wish or intention to flex one's wrist on a particular occasion might 'fit in', understood in the broader context of ongoing processes (bodily, psychological, etc.) associated with continued participation in Libet's laboratory experiment, is far from clear.

[152] This looseness of fit is, arguably, so natural and familiar that even children become adept at exploiting it to their own ends – like the child whose irrefutable reply to their parent's enquiry "Are you tidying your room?" is "Yes, I've just started."

But the more important point is that the significance of the occurrence and timing of such an 'event' is, in itself, far from clear. Certainly, any attempt to straightforwardly identify this (supposedly discrete) event with an hypothesised instant at which we become aware of a conscious intention falls foul of the correctives urged by Strawson and Blackburn. States of conscious intention are not that sort of thing: they are not inwardly-directed states relating 'us'[153] to some internal mental object or content. Rather, these are world- (including body-) directed states that we occupy as whole psychological subjects and agents.

The Libet-type scenario also falls foul of the second corrective I have urged. If actions are not, in general, micro-analytically discrete events occurring at instants in time on some microscopic timescale, then the identity conditions and boundaries of the intentional activity that is choosing to perform a wrist flexing, and the intentional activity that is the wrist flexing itself, were never going to be neatly mapped onto Libet's RP and motor neuron activation measurements without the risk of serious confusion and misinterpretation. The intentional activities are processes extended in time (and space). We need to be able to say something about the relationships between these intentional activities and (i) RPs, (ii) the subjective experiences and the reports of Libet's participants, (iii) the activation of motor neurons that will lead to wrist flexing, and (iv) the overall context of extended activity (intentional and otherwise) of each participant over the entire course of the experiment. But it is safe to assume that neat patterns of identity, or mappings onto neat and discrete mereological sums, will not be in the offing. If Libet (and others) expected such neatness, they could only have done so by mistaking, at the outset, the process-like character of intentional states and activities.

Shaun Gallagher makes a related point in his commentary on Libet's data. Gallagher (2005) insists that we need to be clear about the kinds of actions that are relevant to questions about conscious volition, decision and free will, and that the decontextualised wrist flexing movements of Libet's studies are precisely not of the relevant kind. Deciding on a course of intentional action is, for Gallagher, a process at a level that abstracts from details of bodily movement or implementation:

> Voluntary actions are not about neurons, muscles, body parts, or even movement – all of which play some part in what is happening, and for the most part, nonconsciously – but all such processes are carried along by (and are intentional because of) my decision… to participate in an

[153] Which part of 'us' or 'me' would that be?

> experiment, etc. – that is, by what is best described on a personal level as my intentional action. (Gallagher, 2005, p240)

The primary point that Gallagher is making here is one about the kind of mapping one might expect between voluntary, intentional actions on one hand, and what we might call the components of instances of their bodily implementation on the other hand[154]. As I have been urging above, Gallagher is insisting that voluntary actions are not about, and thus should not be expected to map in some overly neat fashion onto, discrete and particular instances of their bodily implementation, even while any given process or episode of intentional action is "carried along by" our having arrived at a decision, our having occupied a state of intending to act in a certain way.

From this perspective, it makes little sense (despite any appearances to the contrary) to go searching for a discrete event that is our (initial?) moment of conscious intending, as if on the model of some initial billiard ball that would set off the subsequent chain of events. Understood correctly, states of conscious intention and (context-rich) intentional actions just are not sensible candidates for being lined up in a neat causal chain as if they were the ontological equals of things like neurons, muscles, and the various activities of these latter entities.

Bodily Movements and Actions

Gallagher's (2005) take on Libet's data is also helpful because he highlights the extreme artificiality of the candidate intentional actions being performed by Libet's participants. As he succinctly puts it: "The kinds of actions that we freely decide are not the sort of bodily movements described by Libet's experiments" (Gallagher, 2005, p240). This rather obvious but important observation adds a further criticism of the supposedly simple scenario Libet intended, one that goes beyond the correctives we have been considering thus far. Our concern here is not so much with whether or not actions are processes extended in time and space, but with the appropriate characterisation or description of the actions we intend.

Consider the well-worn example, much loved (and arguably abused) by philosophers of mind, of raising one's arm. The most obvious thing to say about the 'action' of raising my arm is that it is not a very good example of an action. The reason for this is very simple: we

[154] Indeed, the section immediately following Gallagher's (2005) discussion of the Libet data is called "Redrawing the map."

don't typically go about our lives engaged in the strange and eccentric activity of intentionally raising our arms. Rather, we go about our lives intentionally reaching up for the book on the top shelf, or indicating to the chair that we would like a chance to speak, or waving a greeting to a friend across the room, or trying to catch the attention of the absent waiter, or making a bid at an auction, or hailing a taxi. These are the intentional activities that fill our lives, and in which arm-raising plays, under appropriate circumstances, a critical role in implementing our intended course of action. In some cases (e.g. the reaching), the arm-raising might be pretty much essential to our unfolding action, given the facts of our embodiment and physical surroundings; in other cases (most of the rest of the previous examples), the significance of arm-raising may be a complex mix of facts (it is easier to be noticed by a friend, waiter or taxi driver of your arm is extended) and norms (our society considers it both normative and acceptable to hail a waiter or taxi by raising your arm, but not by throwing something at them, even though the latter might, as a matter of fact, be more effective). Nevertheless, raising one's arm in each of these cases should not be mistaken for or confused with the intended action that is instantiated, in part, by a token arm-raising. In every case, the appropriate characterisation of the action remains at a level that tends to abstract from the details of its implementation[155].

By implication, wrist flexing is not a good candidate intentional action to serve as a basis for exploring the neuroscience of conscious volition. We no more spend our lives engaging in intentional wrist flexing than we do in intentional arm raising. After a long day hunched over my computer keyboard, I might flex my wrists to combat the stiffness that is setting in to the muscles and tendons in my lower arms and hands. But the intentional act in which I engage here is that of *stretching in order to relieve or reduce stiffness* rather than one of intentionally flexing my wrists. At best, it is only true in an indirect and derivative sense that I intentionally flex my wrists: in the act of stretching my lower arms, I intentionally flex my wrists.

Gallagher (2005) makes the same point while employing a vehicle-content distinction to

[155] We can, as a result, easily track type-identical actions across very different embodied implementations or instantiations of those actions. In the digital era, raising my arm at an auction I am attending in person serves the same purpose as a telephonic bid or an electronic bid submitted in real time over the Internet. All count, in a straightforward and unproblematic way, as implementations of an intention to make a bid in an auction.

highlight the contrast he has in mind:

> I can *say*, in a derivative fashion, that in taking a drink I am freely extending my arm, etc., but only in the same way that I might say that the neurons activated when I see red are the 'red neurons' – I don't mean that the neurons are *actually* red, nor do I mean that the reach and grasp, the muscle extension, the neural activity are freely chosen *per se*. (Gallagher, 2005, p240; italics in original)

So specific bodily movements – wrist flexions, muscle activations and contractions, motor neuron firings – are properly seen as the vehicle to the content that is a particular consciously intended and willed action. As with so many other vehicle-content distinctions, confusion awaits if we do not recognise and respect this distinction between embodied vehicles and intentional volitional contents. Libet's work – or rather, the intended interpretation of the scenario for conscious intentional action that it constructs – does not appear to do this.

These observations about the appropriate characterisation of intentional actions can be used to generate at least two significant difficulties in the interpretation of Libet's data. First, it should by now be clear that participants in Libet's experiments should not be characterised as engaging in individual intentional acts of wrist flexing, whatever the temptation to make this inference given the instructions these participants received and the data generated for 'individual' trials. Instead, the appropriate intentional description of the participants actions should, at the very least, be one that adequately characterises the *ongoing process* of intentional engagement and compliance with the instructions and demands of the experimental scenario as a whole. This process of intentionally and voluntarily engaging with the demands of the experiment, stretching over an extended intermediate time frame on a scale of seconds and minutes, includes individual wrist flexions and motor neuron activations amongst the embodied components or vehicles that implement this intentional plan of action. These embodied components are not, except in a most derivative, artificial and misleading sense, the objects or contents of individual moments of intending.

There is an obvious synergy between this first difficulty and the second general corrective discussed earlier; but the problems they raise for Libet's studies are relatively independent. "Flexing their wrist" is not the appropriate characterisation of the intentional activity of Libet's participants, irrespective of what one may think about the issue of the time frames of intentional actions in general. On the other hand, when the time-frame corrective is combined with the difficulty concerning appropriate intentional description, we find a powerful set of

considerations for doubting the aptness of Libet's experimental scenario as a context for teasing out the complexity of consciously intended and willed action.

The second difficulty I have in mind is less about the appropriate characterisation of the actions performed by Libet's participants, and more about the appropriateness of wrist flexing as a candidate action to consider in the study of conscious intention and volition. Suppose it were claimed, for example, that wrist flexing is in itself a perfectly suitable candidate action to be undertaken with conscious intention, in the same way that looking in the mirror and trying to raise one eyebrow is a perfectly good case of an action engaged in with conscious intent. Libet's experiment throws in some necessary trappings of compliance with instructions, repeated trials, and the additional activities of self-monitoring combined with oscilloscope tracking. But, so the objection goes, consciously-intended wrist flexing remains in itself a perfectly good candidate for an intentional action, in the same way that any instance of conscious voluntary control over the body could, under the right circumstances, properly qualify as the appropriately characterised intentional action of the agent. We might thus still learn something important, and/or be confronted by significant challenges, by studying 'simple' actions like wrist flexions under appropriate laboratory conditions, even if these 'simple' actions are not ones that we typically fill our waking lives with.

It might seem bloody-minded to try and deny that simple bodily movements like wrist flexions or arm raisings or eyebrow raisings could ever be appropriately characterised as *the* intentional action performed by an agent. I do not think it necessary that I make such a denial. Given the first difficulty discussed above, it is enough that the wrist flexing performed by Libet's participants *in his experiments* is *not* appropriately characterised as the intentional action they were engaged in. But I nevertheless think that it is worth emphasising just how far we might need to go in resisting the introduction of simplistic and decontextualised bodily movements as candidate actions that might aid us in understanding conscious intention and volition.

Let us consider a list of relatively simple bodily activities we might engage in with conscious intent: raising one eyebrow, flexing a wrist, winking an eye, letting go of an object we were holding. As we have already seen, each of these should raise some concern in that they have the potential to be used to suggest actions as, in general, discrete events that are pin-pointable in time and space – the caution raised in our second general corrective. Yet, even if we do not

allow this concern to be decisive, there are other reasons to be suspicious of these candidate actions. For a start, we might think that each of them would seem a peculiar candidate for an action unless something more was said about the intentions of the agent. I stand in front of the mirror, and I raise one eyebrow. What am I doing? Evidently, there is a perfectly reasonable sense in which I am just testing out (or perhaps confirming, if I have done this before) what my body can do under my conscious voluntary control. But even this intentional characterisation is richer and more complex than the claim that I am merely (intentionally) raising an eyebrow. Moreover, it is an intentional characterisation of my eyebrow-raising that would distinguish this instance from other instances in which my intention was specifically to see how I looked with one eyebrow raised, or to memorise the feeling of different eyebrow-raisings so as to have better control over these when a mirror is not available, or to exercise my single-eyebrow-raising muscles, or whatever other reasons one could have for standing in front of a mirror raising one eyebrow. And once we have acknowledged these thin additional intentional layers to our simple bodily exertions, we can see that even they take place in a context of ongoing intentional engagement with our interests, projects and the world that blurs their boundaries, including their startings and finishings.

There is an additional problem, though. Decontextualised bodily movements lack meaning and, for this reason, they lack something that is central to the actions we engage in with conscious intent. Isolated instances of wrist flexions, arm raisings, etc. that go *uninterpreted* by the agent and (where relevant) others are not plausible candidates for the study of consciously intended action. Actions are intended and performed under certain aspects or interpretations and not under others. They typically form part of larger patterns and processes of meaning-making and exchange – examples given earlier of just some of the possible meanings with which arm-raising activity can be imbued provide a suitable illustration. A wrist flex that has no significance (beyond experimental compliance) for either agent or observer is thus not a phenomenon worthy of study if our object is to understand the springs of consciously intended and voluntarily willed behaviour.

Conscious Intention

Thus far, I have highlighted a number of reasons for being extremely cautious, if not sceptical, of any simple interpretation of Libet's experimental task and the resulting data. In short, we have good reason to think that Libet's participants' actions are not adequately characterised as mere wrist flexions; and even if they were, such wrist flexions would not be

suitable candidates for studying the neuroscience of conscious intentional action in the simplified scenario set out by Libet.

It might be claimed, however, that this commentary on Libet's work is obfuscatory, in so far as it fails to address a central, singular challenge presented by these studies: how does the timing and sequencing of conscious volitional activity relate and map onto the timed and sequenced neural activity recorded by Libet and company? Perhaps there are valid issues about the fine-grainedness of timescale we can sensibly apply to intentional processes; and perhaps the choice of intentional wrist flexing is not ideal when it comes to the ecological validity of these studies. But for all that, questions about the timing of conscious intentions, in relation to the neural activity associated with preparation for and initiation of these wrist flexions, demand some kind of answer.

My response to this particular complaint has been foreshadowed by my comments on the first general corrective urged by Strawson and Blackburn: we don't so much 'have' conscious intentions as we *occupy states of conscious intending*. When we speak of having a conscious intention, we are describing a state we occupy when we act with conscious intent. The (primarily negative) point made earlier is that the boundaries of such a state, and particularly its beginnings, are likely to be fuzzier than is allowed by the apparently neat logic of Libet's experimental scenario. More specifically, the first moment of awareness of a particular wish to flex one's wrist seems a particularly inappropriate candidate to identify with the beginning, or taking up, of a state of conscious intention, even under the decontextualised conditions for action manufactured by Libet. Moreover, it is an inappropriate candidate because having a conscious intention is not to be in an inwardly-directed introspective state in which we are examining some internal aspect of our psychology[156]. But perhaps these responses could benefit from a little more detail on the view of conscious intention that is being propounded here.

To reiterate, then, 'having' a conscious intention is to be in a state of conscious intent, where this means that we knowingly intend that something be or become the case. Since intentions

[156] Let alone a partly introspective state in which we are trying to track the contents and changes in contents of our conscious awareness *while also* trying to explicitly map certain of these changes onto perceptual input about the position of the dot on the oscilloscope. The point is not to suggest that we are incapable of these feats; rather, it is that this is not an obvious analogue for real-world conscious decision-making and action on the basis of conscious intention.

are for the most part directed at our own actions, it means that being in a state of conscious intent is to knowingly intend to do something. The epistemic element gestures towards the fact that when we intend something, we intend it under an aspect or description, such that knowingly intending something is to be cognisant of, aware of, or to otherwise have access to the aspect or description under which it is intended. In contrast, it is possible for us to intend something, and intend it under an aspect or description, and yet we will be unaware of either the intention or the particular aspect/ description under which something was intended. (A good therapist might, for example, convince us that a behaviour we thought unintentional was in fact intended, or that an act we thought of as intended under one aspect/ description was in fact intended under another, perhaps rather different aspect/ description.)

To act with conscious intent is thus to engage with the world in order to knowingly bring about some state of affairs. When I drive to work in the morning, I do so with the conscious intention of getting myself to work. If anyone were to ask me what I was doing, I would reply that I am driving to work. There are, of course, many other things that I am or might be doing when asked this question: I will be sitting down, I will be holding a steering wheel, I will probably be listening to music, and I will most likely be driving my own car. Different kinds of questions, and different pragmatic contexts, might elicit one or other of these responses from me. But to the generic question "What are you doing?", I would answer that I am driving to work, and this answer would give the clearest indication of my primary conscious intention at the time of interrogation[157].

My consciously intending to drive to work does not necessarily involve me in some ongoing experience or contemplation of a mental entity that is my conscious intention whose (propositional) content is that I drive to work. Nor is it obvious that the (fuzzy-boundaried) beginning of my consciously intending to drive to work should have necessarily involved my experience or introspection of some discrete internal state with this propositional content. It would probably be apt to say that I have a long-standing commitment (and, thus, a reliable background intention) to get myself to work on each and every morning that I am supposed to be at work. If my wife asks me, on a Tuesday evening in term time, whether I intend to drive to work in the morning, I would doubtless answer yes (unless there were some exceptional reason why I would not be doing this). And it is, in part, because I have such a long-standing

[157] There is, of course, nothing to stop me having multiple conscious intentions at the same time, other than perhaps the limits of attention, memory and bodily coordination.

commitment and reliable (background) intention that I would not expect any dramatic internal psychological precursors to my shifting, at some point every morning in term time, from a state of background intention to one of consciously intending to drive to work. To the extent that this shift is a significant one, it might be elicited by an alarm clock, by the thought or memory of an appointment later in the day, by a glance at a clock followed by the thought that it is getting late, or by any other number of internal or external 'triggers'. One or other of these triggers might provoke the (internally or externally expressed) thought "I must get going!". And my knowingly intending to drive to work must include my having the capacity to somehow entertain or represent to myself the relevant states of affairs – my driving, my arrival at work – associated with my state of conscious intention. But, again, nowhere in this perfectly mundane example does there appear the need for a first moment of conscious awareness of the wish or intention to drive to work.

Perhaps it will be countered that the comparison between regular activities – the stuff of habit and settled dispositions – and relatively novel and discrete opportunities for action, such as those presented to Libet's participants, is unfair. Perhaps the operation of conscious intentions in a context of reliable background habits and dispositions need not be punctuated by discrete conscious episodes of intention formation. That does not mean that such episodes should not be expected in contexts of relative novelty where decisions are taken and intentions formed 'in the moment'.

The supposed obviousness of this response is, however, easily exposed as a misrepresentation of the ordinary flow of intentional behaviour. Suppose I engage a stranger in conversation. Since I am conversing with a stranger, I take it that the relative novelty of the situation is beyond question, as is the dynamic 'in the moment' character of the intentional behaviour in which I am engaged. It is my conscious intention to engage this stranger in conversation on the topic at hand; moreover, my contributions to the conversation are offered with equally conscious intent. Yet I do not suppose for a moment that such a conversation is *also* peppered with my formulating and acting on individual conscious intentions for each and every individual speech act I perform. To reiterate the lessons of the preceding sections, this demand for discrete moments of conscious intention formation is misguided. It mistakes the character of mental states like intention, as well as the nature and timeframe of conscious intentional behaviour.

What about conscious control?

Perhaps, however, we ought to anticipate a different kind of objection at this point. Suppose it is conceded that Libet's characterisation of the activities and intentions of his participants is flawed; further, let it be conceded that the critical moment of conscious activity that Libet sought to time and relate to both RP and motor-neuron-activation data is not *the moment* of conscious intention formation. Might it not, nevertheless, be argued that the timing of conscious awareness for a particular wrist flexing, within the context of an ongoing process of engaging with the demands of the experiment, is inadequate to allow much scope for a meaningful and effective *process of conscious control over action*?

The guiding idea behind such an objection could be something similar to that underpinning at least one sort of sceptical response to Libet's conscious veto-based interpretation of his data. Libet has, in various ways, tried to save the notion of free will by claiming that what his participants still demonstrated under his experimental conditions was a variety of 'free won't' – namely, the ability to veto or abstain from a particular wrist flexing at the moment they became aware of the conscious wish to perform the movement.[158] But Libet's interpretation of his own data on the timing of conscious volition suggest an obvious objection to the veto account. Libet's claim is that the conscious wish represents a post-hoc awareness of a process of preparation for action (as represented by RP) that was begun unconsciously. But then why should we not assume that a conscious veto is itself the post-hoc manifestation of unconsciously initiated process of inhibiting or vetoing an action?[159]

Whatever the merits of this criticism of Libet's veto, there is a potentially related point to be made about the time available in which conscious control (whether in the form of a veto, or some other form of conscious intervention) could conceivably be exercised. According to the data, awareness of the wish to flex manifested at between 200 to 150 milliseconds before motor neuron activation. Assuming that a change to, or inhibition of, the eventual motor neuron activation could not be effective beyond -50 milliseconds,[160] this leaves a window of between 100 and 150 milliseconds in conscious which control could be exercised. This small window of opportunity could, of course, be used as partial motivation for the claim that any conscious intervention – veto or otherwise – would most plausibly be seen as the

[158] See Libet (1985, 1999, 2004).
[159] Libet (1999, 2004) addresses this particular question.
[160] See Libet (1999, 2004).

outcome of an unconsciously initiated process on a more generous time scale. But the worry could be taken as having its own merit and implications.

In the light of the second (time-frame-related) corrective proposed earlier in this chapter, we should in general expect that conscious processes, including the conscious initiation and control of intentional behaviour, will unfold over an intermediate term timeframe of seconds and minutes, even hours and days. Taking this very lesson to heart, our imagined respondent might want to claim that the time available for the exercise of any kind of conscious capacity in the experimental situation – veto, guidance control, whatever – is simply inadequate. In winning the battle over whether or not Libet has timed the moment of conscious intending, we lose the war over the potential relevance and efficacy of consciousness in the control of behaviour.

Notice that this sort of argument can be levelled at the likes of Gallagher. In the pages preceding my earlier references to him, Gallagher (2005, p238) asks us to consider an example of him walking along a path, and a snake moving in the grass next to his feet at time T. Between T + 150ms and T + 200ms, and without Gallagher knowing much anything about it, the amygdala in his brain has become activated, triggering a (crude) fear response, and he has jumped away from the snake. Activity of this kind and on this timescale is automatic, reflex-based, and neither needs nor would benefit from conscious awareness and 'control'. On a very different timescale (T + 1000ms onwards), as Gallagher becomes aware of what has happened, and of the stimulus that set off this initial rapid response, a different order of activity emerges. Gallagher asks us to imagine him, at T + 4000ms, recognising the type of snake he has encountered and that it is a harmless variety such that, at T + 5150ms, he moves back towards the snake and (voluntarily) tries to pick it up. This, for Gallagher, is the appropriate timescale of conscious voluntary action, and it is much slower and more temporally extended than that of his initial, automatic response.

There is much to be said for Gallagher's use of this case as a means to highlighting the contrast between the ordinary (intermediate) timeframes over which conscious voluntary action unfolds, on one hand, and the attempt to micro-analyse the timing of action along the lines of Libet's studies. Much of my response to Libet so far is in agreement with Gallagher, and we will revisit aspects of Gallagher's position in Chapter 8. But for present purposes, his claims would seem to lend weight to the current objection: because conscious voluntary

action takes place on scales of seconds, etc., 100 to 200 milliseconds just seems inadequate to the task of giving any meaningful role to consciousness in Libet's trials. Unless we have something more to say about the nature of conscious control over our actions, then we are going to be left with the worry that, at least in Libet-type scenarios, consciousness really does appear to be a mere spectator. And, so the objection would go, once the case is conceded for some cases of simple (if trivial) actions, it is not clear how the rot could be stopped in other cases – including cases like Gallagher's (allegedly) conscious voluntary response to the snake.

I have no immediate response to offer to this objection. Indeed, I think an immediate response would be ill-advised. What is required is a revisiting of issues associated with the control of behaviour. But whereas any previous excursions into this arena were shaped and coloured by the tensions and points of difference between compatibilists and libertarians – most especially including (i) whether indeterminism in volition reduces control over action in freedom- (and agency-) threatening ways; and (ii) whether, as Fisher would put it, compatibilist agents can be said to have regulative control of their behaviour, or merely guidance control – we will now be able to focus on questions about control that are neither prompted nor constrained by these traditional concerns.

In the following chapter, I will thus explicitly turn to questions about the nature of control, and specifically the nature of control in biological systems. The challenge will be to identify the varieties of control on display in nature, and to decide whether one or more of these can be put to work in the defence of a variety of conscious control that deserves to be called free.

Chapter 8

Understanding and Distributing Control

Questions about control are central to many traditional discussions of free agency. Libertarian (especially agent-causal) accounts tend to argue, positively, for a distinctive kind of agent-based control; and yet the apparent loss of control that comes with the insertion of indeterminism into moments of volition is, as I have argued, one of the critical flaws in the libertarian strategy. Even so, Kane tries to exploit a certain kind of reduction in control in explaining how SFAs open up genuine alternative possibilities when choosing between competing and incommensurable courses of action, where we have what he called 'plural voluntary control.' Outside of libertarianism, Smilansky and Fischer both emphasise the importance of less 'meaty' but nevertheless important forms of control – 'local control' (Smilansky) and 'regulative control' (Fischer) – even while they argue that we lack more metaphysically weighty forms of control – what Fischer calls 'guidance control.' And compatibilism in general tends to give considerable importance to questions of control, whether this be negatively through the absence of interference in Frankfurt-style examples (the 'non-intervening controller'), or positively as in Mele's emphasis on self-control. In short, if there is a lack of consensus as to the importance of alternative possibilities to understanding free agency, there is much greater agreement about the importance and centrality of issues and interpretations of control in free agency. That said, there is not much agreement within the traditional debate on what kind/s of control might be needed to secure claims of free agency.

At the conclusion of the previous chapter, we encountered a concern that, whatever the problems and qualifications we might associate with the setup and interpretation of Libet's studies, perhaps the data still show that in the end, there just isn't enough time available (100 - 200ms) for consciousness to make a significant difference to the control of action. On this view, conscious volitional activity has been shown to both *not* lie at the origin of various causal chains leading to action (as per Wegner's account of the illusion of conscious will), as well as *not* offer the prospect of conscious control over the unfolding activity. A sceptical epiphenomenalist generalisation of this position would claim, in line with Wegner's generalisation about the experience of conscious will, that what is revealed about our lack of

conscious control in the impossibly tight timeframe of Libet's experiment is, in fact, a feature of all behaviour: it is both driven *and* controlled by sub-personal, non-conscious processes, with no need or space for conscious guidance and control. Libet's data, like the carefully controlled automaticity studies, just helps bring the nature and extent of such sub-personal, non-conscious control into focus. Consciousness might keep us (imperfectly) informed about what we are up to, and perhaps even deliver up a user illusion of some kind[161], but we are misled if we think that consciousness does any more than this by way of active intervention and control of our behaviour.

In short, what is it stake here is the idea that we have some kind of *conscious executive control* over our lives as agent – the kind of conscious executive control whose absence (as discussed in Chapter 5) represents the threat that we are merely, or even just mostly, agent automatons (AAs). The burden of the current chapter is to tackle this concern head-on through a sustained examination of ideas and models of control. I will argue that, while we might be inclined to make certain common sense (and philosophical) assumptions about the nature of control implicated in human agency, the form and distribution of control in complex biological systems, including our own bodies, is a much more complicated affair that requires empirical and theoretical illumination, as well as careful philosophical theorising. A defence of free agency may well require a characterisation and defence of distinctive varieties of human executive control, especially involving consciousness, but it need not be just one image, or *one* kind of control that is implicated; nor need it be the case that distinctively human varieties of control over behaviour always amount to more than a matter of degree, at least when compared to other conscious animals (although perhaps not to non-conscious artifacts).

Assumptions about Control and Executive Control

As intimated in Chapter 5, the image that is most clearly threatened by the spectre of the AA, and thus the image that is most obviously threatened by the evidence and arguments presented in Chapter 6, is that of the conscious agent, in control of themselves and their behaviour. This is not an image of the agent as possessed of some strong form of self-control (as understood by, for example, Mele (1995, 2006) – see Chapter 3 and Chapter 5 of this thesis) conceived in strongly cognitive, deliberative and rational terms. It is, first and

[161] See Dennett (1991) and Nørretranders (1998) for more on the idea of user illusions.

foremost, an image of the agent and their relationship to their motives, plans and intentions, as well as to their actions. In this familiar image, the human agent has a significant degree of what we might call *conscious executive control* over their actions, accompanied by a fallible but (mostly) reliable conscious awareness of their plans and intentions in action.

In Chapter 7, I began to clarify part of what this might mean, at least in relation to what it means to act with conscious intention. Conscious intention is not an inwardly-directed state involving a relation to some mental item. Instead, having a conscious intention is to occupy a state of conscious intent, and to act with conscious intent is to engage with the world in order to knowingly bring about some state of affairs. Given this clarification, the latter part of the image just described (i.e. the part involving conscious awareness) fits nicely with a careful understanding of conscious intention and action; but it is not yet clear if the assumption regarding executive control is entirely apt, even if acting with conscious intent presumably must require a significant role for conscious control of some variety.

What could this assumption of executive control amount to? One tempting answer might be that, if we strip the word 'ultimate' of its libertarian and incompatibilist connotations[162], we could say that we think of our conscious selves as, much or most of the time, being the *ultimate arbiter* of what we do. In contrast to libertarian notions of ultimacy, there is no necessary emphasis here on being the ultimate *source* or initiator of what we do. Being the ultimate arbiter of (much or most of) what we do relates as much to what we *allow* ourselves to do as it does to what we choose to do, where our choosing is part of some process that initiates a course of action. And, as a first pass, the *ultimacy* of such control can be cashed out (again in a manner that avoids metaphysical complications of the traditional debate) in terms of an ability to consciously monitor, modulate, modify, suspend, persist with, and terminate behaviour in the light of our conscious intent and the monitored outcomes of our ongoing activity.

This first gloss on the notion of conscious executive control suggests, in turn, that part of what is so distinctive about conscious agency is a capacity for centralised, instructive control. This control is *centralised* in so far as there is some functional equivalent of a single, central control room in which information about the state and activities of the system is brought

[162] As in Kane's (1996, 2002a) notion of Ultimate Responsibility.

together, and from which executive instructions to the rest of the system are issued. The form of control is *instructive* in so far as it is based on (functionally[163]) centralised planning, such that the executive instructions that are issued to the rest of the system derive from the central plan, and thus convey information about that plan. Similarly, if we are to understand the monitoring aspects of executive control along these lines, we would expect to find that information about the state and activity of the system is gathered centrally, such that monitoring comprises a comparison between the central plan and the actual progress/ outcomes so far, with these feedback mechanisms guiding any requisite modifications of the central plan, ongoing executive instructions, or both.

This picture or model of conscious executive control should strike most of us as being familiar, intuitive, and lacking in any obvious philosophical baggage. It is, arguably, just the kind of image of agency presupposed by the basic logic of Libet's experiments. And a model of just this kind is helpfully described by Francois Schroeter (2004), under the label 'basic executive control', in his critique of action theory and endorsement-based accounts of autonomy. While the details of Schroeter's actual project are not central to my concerns at this point, his account of basic executive control is useful for expository purposes.

In the midst of his critique of what he calls 'standard action theory', and with it also endorsement-based accounts of autonomy, Schroeter (2004) presents us with an account of what he thinks basic executive control amounts to for human agents. In particular, he highlights three aspects of such basic executive control: (i) the agent that is best identified as the system wielding basic executive control is the *conscious self*; (ii) human actions of a relevant order of complexity are carried out in accordance with action plans, and because we have these action plans, basic executive control is characterised by the agent (i.e. the conscious self) being in a position of, in general, *knowing what they are doing*; and (iii) action can be *initiated by central commands* – "the motor system can be directly activated by a command issued by the conscious self" (Schroeter, 2004, p645).

Of course, much more needs to be said about such a view of agency and executive control if it is to stand up to both intuition and a wide range of relevant empirical and conceptual considerations. But for expository purposes, Schroeter's account will do nicely. None of (i) to

[163] The emphasis on functional centralisation is so as to avoid any unnecessary assumptions about physical centralization (in the sense of localising of function).

(iii) sound especially outrageous *prima facie*, or excessively loaded with philosophical baggage. One key idea under (ii) seems particularly uncontroversial, especially in the light of what I have argued in Chapter 7 about the link between acting with conscious intention and our being aware of our intentions in action, at least under some relevant description. In this sense at least, conscious human agency is indeed marked by us, in general, knowing what we are doing. The central claim under (i) will raise various hackles on the part of sceptics about the self, consciousness, or both; but I think that Schroeter is correct in claiming that this is our common sense assumption about agency – pre-philosophically, we identify ourselves as agents with some notion of our conscious selves. And claim (iii) is not some metaphysically overburdened claim about the nature and possibility of agent-causation (*a la* agent-causal libertarianism), but rather an expression of the common sense idea that, at least some of the time, we (*qua* conscious selves) can issue commands or instructions for our selves (especially our bodies) to do things that we (*qua* conscious selves) have decided we will do.

As is so often the case, the trouble with intuitively appealing, common sense characterisations is that they so easily land us in trouble. Each of (i), (ii) and (iii) is under threat, in some way, given the evidence of Chapter 6 regarding automaticity, readiness potentials and the apparent illusions of conscious willing. The claims regarding automaticity advanced by Bargh *et al* directly challenge (i) and (ii), while suggesting that the scope for (iii) might be vanishingly small. Wegner rejects both (i) and (iii), while his view of (ii) is less clear and, at the very least, counterintuitive. For Wegner, we may in one sense generally know what we are doing, but this is complicated by the claim that, in another sense, we are *always* mistaken about what we are doing because we suffer from the illusion that our actions are a product of our conscious willings, as per (iii). Libet's data need not prescribe any particular position on (i) and (ii) – certainly, Libet's own interpretation of his data would suggest that he endorses versions of (i) and (ii). His data prompt puzzles about (iii), however; and, given these puzzles, the data also raise questions about the nature and possibility of executive control wielded by the conscious self (i) in the light of centrally-formulated action plans (ii).

In summary, it is relatively easy to make a case for there being a particular, distinctive notion of agent control over themselves and their actions, and to further unpack this notion of control by way of ideas of a conscious, centralised, self-aware and instruction-issuing form of basic executive control that we attribute to the conscious self. Yet it seems that it is just such

a picture of basic executive control that is most obviously threatened by the evidence and arguments of Chapter 6. How, then, should we advance the immediate objective of understanding control in human agents in such a way as to better defend our claims to be free agents?

Centralised Thinking

A sensible place to start is by asking whether distinctively human executive control, such as we might think is implicated in free agency, really needs to be centralised and unified in the way that seems to be implied by Schroeter. That is, there is value in asking whether notions of conscious executive control in human agents need to be strongly tied to ideas of a centralised controller that is unified and identified with the conscious self – a central planner and instruction issuer.

One reason to be cautious in taking this route to explicating the notion of conscious executive control captured in a warning sounded by Andy Clark (1997). Clark draws on MIT researcher Mitchel Resnick's (1994, 1996) idea of 'centralized thinking' in warning against our tendency to look for single-factor explanations of complex phenomena:

> People seem to have a strong preference for centralization in almost everything they think and do. People tend to look for *the* cause, *the* reason, *the* driving force, *the* deciding factor. When people observe patterns and structures in the world (for example, the flocking of birds or the foraging patterns of ants), they often assume centralized causes where none exist. And when people try to create patterns or structure in the world (for example, new organizations or new machines), they often impose centralized control when none is needed. (Resnick, 1994, p120; italics in original)

The tendency is not just epistemic – it is not just a question of what we look for, and what we assume is or must be the case, in our various attempts at describing and explaining various patterns we encounter in the world. Centralised control exists, and in many of the places where we will find it, it is there because humans put it there. Command economies and serial computers designed around a single central processing unit would, I take it, be some of the relevant examples that Resnick and Clark might have in mind.

As self-aware and self-reflective agents, our own agency and how we understand that agency presents an interesting case of the potential for, and potential consequences of, centralised thinking: (a) we may or may not experience ourselves, pre-reflectively, as agents who exert

the kind of centralised executive control that Schroeter describes[164]; (b) we may well be inclined to characterise our agency in centralised terms when we reflect on our experience and engage in various tasks of explanation and justification for our behaviour; (c) we may well be inclined towards centralised thinking when we engage in various third-person explanatory projects involving other human agents. Such explanatory endeavours would range from the mundane 'folk' activities of explaining and predicting the intentional activities of our fellows, to more specialised and technical explanatory efforts within psychology or economics; and (d) depending on what we experience, think and come to believe under (a) through (c), we may aspire to and strive for a form of centralised agency that conforms with aspects of our experience and beliefs, but that may or may not fit well with our real character as embodied agents.

We will touch on these issues repeatedly in what follows. Clark's (1997) immediate purpose in highlighting our tendency towards centralised thinking is twofold. First, he cites Resnick in the context of discussing a view of human development that is, as Clark (1997) sees it, one that does not require 'blueprints'. Second, Clark (1997) is explicitly addressing questions about control, and in this context argues for the importance of *decentralised* control as an alternative to exclusive, even excessive, assumptions about the need for centralised control.

Clark thinks of blueprints as, essentially, centralised plans, and his primary claim is that, at least in the case of a range of behaviours that have been the focus of various developmental studies[165], "there is no 'blueprint' for [this] behavior in the brain, or in the genes" (Clark, 1997, p40). There are two ideas of blueprints at work here. One is the idea of a central plan – probably 'coded' in the genes, but also perhaps 'transcribed' somewhere else such as in the brain – according to which development unfolds. Such a blueprint would dictate and guide the course of the developmental process itself. The second is the idea of a centralised plan for specific behaviours, such as those that have been the subject of inquiry in developmental studies. These include walking, reaching, and learning to cope with slopes. In this context, the absence of a blueprint for behaviour would imply the absence of a single, central plan for behaviour – most likely interpreted as a centrally located *motor plan* of some kind in the

[164] Schroeter clearly assumes we do have this experience; it is arguably the case that Libet and (even) Wegner also assume this to be our experience.
[165] Clark (1997) draws almost exclusively on the influential work of Thelen and Smith (1994). Aspects of this work are discussed later in this chapter.

brain. Similar to a developmental blueprint, such a motor plan would dictate and guide the unfolding of a behavioural sequence, such as reaching for a desired object.

Turning to issues of control *per se*, Clark (1997) is intent to outline and highlight various possibilities for decentralised control in biological (and other) systems as a kind of corrective against the tendency towards centralised thinking. Whereas the absence of a blueprint suggests the absence of a centralised plan, decentralised control in a system highlights the absence, or potential redundancy, of a central controller in a system that manifests coordinated, adaptive behaviour. For Clark, adequate reflection on the existence of systems exhibiting decentralised control, along with recognition of the robustness and highly adaptive character that such systems can display, should make us cautious in asserting too quickly, along with Schroeter, just what form basic executive control must take in human agents.

In order to see these ideas put to work, it will be helpful to consider a detailed example of the kind of processes Clark has in mind and, through this, the light that might be shed on issues of control over behaviour in human agency. As it happens, we have already met with such an example in earlier discussions – the case of children learning to reach.

Bodily and Mental Flailings

In Chapter 3, we saw that Mele (1995) draws an analogy between how a child progressively gains control of their bodies and how they gain ever greater control of their minds and wills:

> Our earliest limb movements are mere flailings, but the flailings themselves play a role in our gaining control over the motions of our limbs. Our earliest sequences of practical thoughts might best be viewed as mental flailings, flailings that play a role in our becoming competent practical reasoners. Eventually, we are able to represent options, to select among them, and to take steps toward the selected goal. By that time, we are beyond mere bodily and mental flailing, and we are able to *choose*. (Mele, 1995, p.228; italics in original)

Mele's (1995) analogy is both apt and misleading. It is apt in so far as it suggests some kind of continuity of issues between our gaining control over our bodies and our becoming competent practical reasoners. It suggests a potentially important relationship between bodily 'flailings' and mental 'flailings' in the emergence of distinctively human forms of control we most closely associate with agency and choice. It is misleading, however, in so far as one might try to read too much into the *analogy* itself – that is, one might be tempted to infer that, instead of a continuity of issues, the analogy instead implies a clear and sustainable distinction between issues of bodily control, on one hand, and issues of mental control and

practical reasoning on the other.[166] Certainly, the research and theory that Clark (1997) discusses suggest that bodily control is not so much *analogous to* or *like* the acquisition of mental control – for the most part, these are part and parcel of the *same* process. Indeed, it may be that 'flailings' (which hardly seem like much of a model of or basis for control) are quite generally crucial to the development and emergence of control in human agents, including any relatively distinctive forms of executive control we might lay claim to.

Let us, then, look in greater detail at the research that Clark finds so thought provoking, and so challenging to our ideas about centralised plans and centralised control. Clark (1997) is primarily interested, in this context, in the developmental research of Esther Thelen and Linda Smith (Thelen & Smith, 1994). Their research on the development of reaching and grasping behaviour in infants, as summarised by Clark, resonates particularly strongly with the quotation from Mele above. In the extract below, Clark describes two of the infants whose longitudinal development of the ability to reach formed part of a larger study (four infants in total) originally reported in Thelen *et al*. (1993), and discussed at length in Thelen and Smith (1994):

> One infant, Gabriel, was very active by nature, generating fast flapping motions with his arms. For him, the task was to convert the flapping motions into directed reaching. To do so, he needed to learn to contract muscles once the arm was in the vicinity of a target so as to dampen the flapping and allow proper contact.
>
> Hannah, in contrast, was motorically quiescent. Such movements as she did produce exhibited low hand speeds and low torque. Her problem was not to control flapping, but to generate enough lift to overcome gravity. (Clark, 1997. p.44)

In the context of Thelen and Smith's (1994) work, Hannah and Gabriel provide a wonderful contrast that helps undermine certain assumptions about development. Both infants are learning to reach. As an abstract developmental goal, they have this much in common. But these infants have to solve *different problems* in order to achieve this overarching developmental goal. This neatly illustrates the ideas, mentioned earlier, about development without blueprints. Hannah and Gabriel's reaching behaviour does not represent the unfolding of some common developmental plan. Hannah has one set of problems to solve – problems about generating activity, developing enough torque to get her arms and hands moving in the right direction – while Gabriel faces an almost entirely different set of challenges – shaping, constraining, dampening his energetic arm movements to as to turn these into something that approximates a controlled reach and grasp of an object. This

[166] It is not clear to me to what extent Mele (1995) intends his example to be read as an analogy. The first three sentences are consistent with an analogical reading. The last sentence, by placing bodily and mental 'flailings' alongside each other, is suggestive of greater continuity with respect to the issues involved.

variability in the developmental problems faced, and in the solutions that need to be generated, by these children suggests both that there is, in fact, no single developmental 'plan' unfolding in their cases, and that at any rate there could be no single plan that would cater for the variable challenges involved.

The development of reaching, as evidenced in these infants, also provides a useful case to illustrate the notion of what Clark (1997), again following Thelen and Smith (1994), calls *soft assembly*. Rather than following some single, centralised plan for the unfolding development of reaching behaviour, Hannah and Gabriel can be thought of as soft assembling solutions to problems that they face in their increasingly intentional and intentionally-driven engagements with the world, such as wanting to reach out and grasp an object like a toy. For Clark, the idea of soft assembly is both critical to an appropriate understanding of development, but also essential to an appropriate understanding of the achievement of control in various systems, including little systems like Hannah and Gabriel.

Soft Assembly

The notion of soft assembly is, in a critical sense, contrastive. Design and developmental processes involving soft assembly are to be understood in part through a contrast with what Clark (1997) calls 'hard assembly', and at the root of this contrast lies a distinction between different forms of guidance and control. In brief, hard assembly is characteristic of systems (including many systems designed by humans) that have a strong centralised controller issuing commands to the system's parts based on 'internal' representations, blueprints, specifications, centralised computations and/or instruction sets. The centralised controller can command a given range of movements or operations of the system's parts, and does so on the basis of 'plans' for how a specified goal might be achieved by following a definite sequence of operations. Critical to the idea of a hard assembled system, then, is a centralised controller possessed of 'knowledge' about the systems parts and potential operations, capable of performing relevant computations that relate these to the attainment of a given goal, able to issue commands to the rest of the system that initiate and control movement towards successful achievement of the goal, and able to receive and accommodate to feedback from the system (and the world) as a given performance unfolds. In terms of knowledge, control and guidance, hard assembled systems are heavily asymmetric, with the centralised controller being critical to just about all aspects of system behaviour.

In contrast, soft assembled systems tend to be *decentralised*, piecing together coordinated goal-directed behaviour out of the dynamic interplay of any number of variables and constraints both inside and outside of the system. In the place of a centralised, instruction-issuing controller, we find something more like the emergence of coordinated goal-directed behaviour through the dynamic interaction of internal and external characteristics, capacities and forces. In terms of symmetry, Clark (1997) sees soft assembled systems as displaying something more like an 'equal partners' approach with respect to the roles played by the various interacting factors and forces, such that an exclusive or excessive emphasis on any one factor (*the* cause, *the* deciding factor, as Resnick put it) is likely to lead to fundamental misunderstandings:

> To focus on any one of these parameters in isolation is to miss the true explanation of developmental change, which consists in understanding the interplay of forces in a way that eliminates the need to posit any single controlling factor. (Clark, 1997. p.42)

Besides the potential for misunderstanding, the move away from a single, centralised controller also has important consequences for resources and adaptability. If no instruction-issuing central controller is needed, then the centralised 'knowledge' bases, instruction sets, computational resources, and command and monitoring capacities required for coordinated goal-directed behaviour are either not needed, or (more likely) are of a very different kind from what we would expect to find (or posit) in the hard-assembled case.

Clark (1997) provides a helpful illustration of the contrast between hard- and soft-assembled systems – this time drawing on the work of Pattie Maes of MIT – by describing a case of a scheduling system in which we want to allocate jobs to a group of machines in order to maximise on efficiency and productivity. I like to concretise the abstract description provided by Clark by imagining a group of networked photocopying machines. Our task as system designers is to come up with the most efficient *and* robust system for distributing individual print jobs amongst these machines, given their various capabilities, service cycles, faults, and their existing cues of print jobs. A centralised, hard-assembled solution to this challenge would be to have one part of the system – or, as may be the case, a separate system – act as central monitor and job allocator. We might imagine, for example, a computer interface hooked up to a scanner, at which a user can enter the details of a given print job and scan the relevant pages to be copied. The hard-assembled software system controlling the larger system would then compute, relative to its latest information about the status of each machine in the network, the most efficient strategy for allocating the new job. This setup requires that

the central controller be in constant contact with each production unit in the system, and that it receive accurate information about each of these units on at least all of the variables mentioned earlier – print capabilities, maintenance schedule, faults, existing job list, etc. Furthermore, it must constantly update all this information, and reschedule (i.e. modify its central plan) based on ongoing updates from the network.

By contrast, a decentralised solution to this challenge of task scheduling, drawing on ideas of soft assembly, would have a network of machines without any central controller to act as either monitor or allocator of individual jobs. Imagine, instead, a network of printing machines each of varying capabilities, maintenance status, etc., that have been hooked up such that (i) any individual machine can be used to scan and programme the required specifications for any new print job; and (ii) every machine in the network can make a simple calculation of estimated time to complete a given print job, given its own local state, and submit this 'bid' to the machine at which a new job has been submitted. The machine in the network with the fastest estimated completion time 'wins' the new job, and the relevant data gets sent across the network to that machine. End of story.

Notice the numerous important differences in the soft-assembled, decentralised solution to the task-scheduling problem when contrasted with the hard-assembled centralised alternative. Most notably, there is no central controller in the system tasked with monitoring the state of all the system parts, and with making costly calculations of job distributions based on information about the system's state that is, in turn, being constantly updated. Indeed, there is no part of the system that has information about the system as a whole – this information is distributed throughout the system in a properly decentralised fashion.

Not only does this setup obviate the need for complex software capable of receiving and integrating system status information and performing system-wide scheduling calculations on this basis; it also makes the system immune to the possibility of catastrophic failure that would result from a breakdown in the central monitoring and allocation system. The system is, in this sense, robust in ways that a hard-assembled centralised version is not. The soft assembled system is robust in other ways too, most notably because the local failure of any given machine has minimal effect on the operation of the system as a whole, even if overall productivity is negatively affected for the period the machine is out of action. All that happens is that the machine that fails, or that requires down-time for maintenance, does not

submit bids within the system over that period. The rest of the system carries on as before. There is no need to recalibrate or recompute distributions and distribution algorithms for the system, because there are no centralised systems of these operations.

The efficiency, adaptability and robustness of this soft-assembled decentralised system comes from a combination of relatively simple local properties for units in the system (specifically, each unit being able to compute and submit a bid based on its own state) with important system level properties (interconnectedness of units, the allocation-to-best-bid principle) that result in both information about and control of the system being distributed and non-local. In this sense, the capacity of the system to efficiently schedule tasks is properly an *emergent* feature of the system; and it is specifically *not* the result of an explicit instruction set constructed for the purpose of computing task schedules in a system of given parameters, and to be implemented by a centralised monitoring and controlling executive system.

This extended illustration of the contrast between hard- and soft-assembled systems brings into focus not only the potential power and utility of systems manifesting decentralised control, but also the potential overlap of ideas about soft assembly and decentralised control with ideas about self-organising phenomena, including self-organising behaviour in various biological systems.

Self-organising behaviour and systems: a modern take on emergence

Clark (1997) himself makes an explicit link between his claims about soft assembly and more general notions of self-organising behaviour in a variety of systems, including flocks of birds and colonies of foraging ants. Resnick (1994, 1996) similarly makes use of these examples, along with traffic jams and slime molds. This link, in turn, points towards a contemporary vein of thinking about the concept of *emergence*, where the latter term has been stripped of most of its historical associations with vitalism and has, instead, been put to work in describing and explaining systems that evidence emergent complex and coordinated behaviour in the apparent absence of any central controller (a dictatorial queen ant in the case of an ant colony; any semblance of a nervous system in a slime mold – which is in fact a collective of individual organisms) by 'following' limited sets of relatively simple rules.[167]

[167] For discussions of non-vitalist emergence (outside of mainstream analytic philosophy), see, for example, Johnson (2001) and Holland (1998).

A description of the behaviour of ant colonies through the lens of self-organising behaviour[168], for example, suggests the emergence of system-level coordination and control (as if instructed by a dictatorial queen ant and/or her 'generals') that is in fact built on elements of the system (individual worker ants) responding to relatively simple chemical cues that cannot in isolation be viewed as constituting information about the colony's collective goal. Empirical evidence suggests that individual ants have the capacity to respond differentially to a variety of chemical (pheromonal) cues, and to gradients of such chemicals. Interpreted from our point of view, we might attach labels to some of these cues such as 'I am carrying food' or 'Run away' (alarm). And yet, understood strictly as an emergent system, it is far from clear to what extent any kind of 'information transfer' is needed in detecting and responding to these signals[169]. Individual ants need to detect and respond differentially to chemical signals, not 'read' or 'interpret' them.

Let us put some illustrative flesh on these conceptual bones. Imagine a colony of ants whose foragers set out for the day. There is evidence[170] to suggest that exploratory foraging is largely based on the more-or-less random explorations of individual ants. Yet most readers would probably think of prototypical signs of ant foraging behaviour as involving a busy column of ants blazing a trail between their nest and a food-source in a disciplined, almost military fashion, as if under the control of some central command structure. The appearance of central coordination, command and control is, however, an illusion. Instead, these columns of coordinated food-retrieval activity *emerge* from a much simpler pattern of foraging behaviour and chemical trailblazing.

Random foraging behaviour will, on average, reliably result in individual ants encountering a food source. When an ant 'shoulders' their load of food and returns to the nest, all they need do is leave a particular chemical trail that other ants can detect – the release of the relevant pheromone being a simple, reliable, 'blind' response to the presence of food. Now, if another random forager encounters this chemical trail, all they need have (in order for the coordinated

[168] See, for example, Johnson (2001). Clark (1997) also provides a brief account of ant foraging behaviour that is largely consistent with the one I offer here. The precise details of the mechanisms used by various ant species do not matter so much as the idea of how there foraging (and other) activities can emerge from following relatively simple, local-level 'rules' or dispositions.
[169] I qualify this statement in this way in part because Johnson (2001), and many of the theorists and researchers he discusses, display a tendency to talk about information and even 'semiochemicals' (implying some kind of semiotic capacity).
[170] See Johnson (2001).

retrieval system to get off the ground) is a selective response or bias to follow that trail in the direction of decreasing chemical gradient, the trail being strongest where the food-carrying ant was most recently – namely, closer to the nest. If they encounter a trail and turn in the direction of decreasing chemical gradient, they will on average eventually arrive at the same food source encountered by the first ant.

As more and more ants follow the same simple pattern, the trail of 'food carrying' chemicals will grow stronger, increasing the likelihood of other foragers successfully detecting and following the trail to the food source. Moreover, any ants emerging from the nest itself will be increasingly likely to encounter the trail before even engaging in any random foraging. Thus can we explain the elegant emergence of a column of hard-working foragers moving determinedly to and from the nest. And, once the food source has been exhausted, the 'food-carrying' chemical trail will naturally dissipate, resulting in a gradual return to either the nest or to random foraging.

The positive lesson to be extracted from this and other examples of self-organising behaviour (birds flocking, slime molds, etc.) is that you can get emergent, coordinated, system-level behaviour from the action and interaction of relatively simple units within the system that each behave according to relatively simple 'rules' (or, better, dispositions) which involve relatively simple differential responses to a relatively simple set of local conditions or variables. The negative (or cautionary) lesson to be extracted from cases of self-organising behaviour is that you do not need to assume some form of centralised control and coordination (including attendant systems of information gathering and exchange) whenever a system is encountered that displays interesting, complex and coordinated system-level behaviour.

Beyond Self-organisation: Instruction versus Selection

So far, we have encountered illustrative examples of soft-assembled systems and, more generally, systems displaying self-organising behaviour, that involve the interaction of relatively independent units to produce an emergent level of system functioning and coordination. Clark (1997) clearly thinks that soft-assembly-based accounts of human development and behaviour will follow along similar lines to these other explanations that utilise ideas of self-organisation. But we have not yet seen a case where these principles have been applied within a system that is itself, for various descriptive and explanatory purposes,

an obviously integrated whole – like an individual organism. Before we make too much of the prospects for decentralised control over behaviour in creatures like ourselves, it would be helpful to ask about the potential relevance of notions of self-organisation *within* biological systems that appear to have greater systemic unity and integration than ant colonies, slime molds, flocks of birds and ecosystems.

A fruitful avenue to explore in this regard involves a subtle shift of emphasis. Consider, once again, the case of our foraging ants. I said earlier that we might be tempted to impose labels on the relevant chemical trails involved in the emergence of coordinated foraging activity, such as 'I'm carrying food' for the trail left by an ant returning with a load to the nest. I also noted, however, that whatever the temptation to read information into these chemical trails ('semiochemicals' – see my earlier footnote in this chapter), there was no need to assume information transfer regarding either the 'meaning' of the chemical trail, or the state and goals of the system as a whole, in order for the system-level behaviour to emerge.

This emphasis on information and information transfer is at the centre of the contrast that Gerald Edelman puts to work in his approach to immunology, neuroscience, and biology in general. Edelman has, for many years, grounded his theories and hypotheses on a critical contrast between what he calls *instructive* models and *selectional* models. Both the immune system and the brain are, for Edelman, systems of *recognition* charged with the task of responding to variable (and novel) 'inputs' in a complex and coordinated fashion that serves the ends of the organisms in which these systems are located. In both cases, Edelman holds that the key to understanding how these systems work lies in selectionist thinking, as contrasted with instructionist thinking. The contrast is most clearly (and, arguably, neutrally) illustrated in the case of the immune system.[171]

In his book *Bright Air, Brilliant Fire*, Edelman (1992) gives perhaps his most accessible description of the historical developments in thinking about the immune system that led from an instructive model to a selective model. According to Edelman (1992), the theory of instruction that prevailed prior to the influential work of Burnet in the late 1950s had involved the idea that the antibody molecules produced by lymphocytes would bind and *mould* or *fit* themselves to the shape of antigens encountered in the body. Having thus 'read'

[171] Edelman's Nobel Prize was awarded on the basis of his work on the immune system.

the shape of the antigen before detaching themselves from it, the antibodies would then keep the shape they had taken on when binding to the invader. By means of this transfer of information about antigen shape, the system was then left prepared to meet and bind to any molecules of the same kind in the future.

Edelman frames this as what he wants to call an *instructive* process because of the posited transfer of information about antigen shape that is hypothesised as taking place when the foreign molecule is first encountered. The shape is 'remembered' by way of the antibody's holding the complementary shape it has taken on in the binding process for use in future acts of 'recognising' the same kind of antigen on the basis of shape and fit[172]. The system is *instructed*[173], through information transfer, about to the shape of a given antigen, and the system is then left with a *blueprint* for dealing with future encounters involving that type of antigen.

According to Edelman's (1992) historical account, the rival theory put forward by Burnet (and further developed and tested by Edelman) proposed a *selectionist* model, based on the hypothesised existence of a population of pre-existing antibody variants. On this selectionist model, the immune system is characterised by a built-in capacity to produce antibodies with a vast range of shapes (in terms of their binding sites). Any previously unencountered foreign molecule entering the body is thus not confronted by some generic antibody capable of somehow fitting itself to the antigen's particular shape (instruction). Instead, it encounters a population of immune cells with a large repertoire of existing antibodies with an already diverse range of shapes, one or more of which is likely to have a shape that is sufficiently complimentary to bind with the novel invader. Successful binding then triggers the lymphocyte to start dividing, thus giving rise to a proportionately greater number lymphocytes in the system as a whole that carry the 'successful' antibody/ies able to bind with the specific antigen that was encountered.

The immune system thus acquires the ability to deal with a range of pathogens by way of a selective process involving differential levels of lymphocyte reproduction from amongst a pre-existing population of variants. Because successful binding sets off a process of cell

[172] The scare quotes simply indicate my caution in applying these psychological terms to these obviously non-psychological and sub-personal processes.
[173] 'Programmed' could be substituted for 'instructed', so long as the metaphorical implications of this are kept in mind.

division, successful binders become successful replicators whose distribution in the population of lymphocytes increases as a result, leaving the organism adaptively prepared for future encounters with the antigens it has encountered.

On this basis, Edelman (1992) views the immune system as a kind of *recognition system* with a form of *memory* embodied at a cellular level. At the same time, the selectional basis of the system, grounded in pre-existing variation in the population of antibodies[174], also leaves the organism ready to adaptively respond to new invaders, including molecules never before encountered by the organism or its ancestors. We have a kind of recognition system with a memory, built up over the individual life-span of the organism, capable of largely adaptive responses to both previously encountered and novel pathogens, *without* requiring a mechanism of instruction (via information transfer regarding shape) or centralised control (instructing the system as to which antibodies it should reproduce in the face of a particular antigen).

The immune system, understood as a selectionist system, provides a clear example of what is effectively a subsystem within a larger system – the organism – in which coordinated, adaptive responses that serve the purposes and welfare of the larger system can be achieved without the need for centralised instructive control. Since selectionist systems represent examples of soft-assembled solutions to adaptive problems based on mechanisms of decentralised control, we can expect to find soft assembly and decentralised control in any biological system that displays such selectionist characteristics. And we find such additional evidence for the relevance of self-organisation and soft assembly to human behaviour and control in certain influential accounts of brain development, including Edelman's overarching theory of neural development – the Theory of Neuronal Group Selection (TNGS)[175] – as well and in the work of Terrence Deacon (1997) on the neurodevelopmental process he calls displacement. Moreover, there is evidence to suggest that our tendency towards centralised thinking shows its head when we think about bodily development more generally, especially

[174] Edelman (1992, pp77-8) provides a brief but clear account of the genetic mechanisms that assist in generating such variation. In essence, his research suggests that antibody production by the lymphocytes involves the production of polypeptide chains that have both stable and variable regions, the variable regions being associated with the binding site for the antibody. Variations in the structure (and, hence, the shape) of the binding sites appear to be enabled by a kind of 'jumbling' of parts of the genetic material in the lymphocyte, this taking place over the course of the individual organism's life-span, and in such a way as to make each individual's repertoire of antibodies unique to some extent.

[175] A theory endorsed, for the most part, by Thelen and Smith (1994).

in relation to ideas about genes, genetic 'blueprints' and genetic expression. It is thus worth our while visiting these conceptual and empirical arenas to see how the tension between centralised thinking and these rival ideas plays out.

Brain-body Development

Given our alleged tendency to look for or assume the existence of centralised controllers within complex systems[176], it is not surprising to find a strong tendency to view embryonic development in general, as well as the specific case of neural development, as a story about controllers – genes – somehow being involved in issuing orders and instructions for the building of an organism according to the genetic blueprint provided in the individual's DNA. While self-confessed genetic determinists are admittedly thin on the ground[177], the tendency to view development as a progressive unfolding of a series of instructions for building a body, with these instructions being somehow laid down and encoded in an organism's DNA, is a widespread and familiar feature of developmental discussions across multiple disciplines. And yet a range of theorists (including Deacon, 1997; Edelman, 1988, 1991; and Rose, 1997, to mention only a few) have provided detailed and convincing insights into the developmental process – most particularly the development of the brain and nervous system – that go a long way towards undermining any such gene-centric interpretations, with their attendant assumptions about genetic control and (centralised?) orchestration of ontogenesis. Instead, these theorist put various distributed, self-organising and selectional (as opposed to instructional) dynamics and processes at the heart of the developmental story.

On the views offered by these theorists, embryonic development is a dynamically interactive process in which the gametes, the embryonic environment, the relative positions of cells within the growing embryo, and irreversible developmental events (such as cell death) all have a vital role to play alongside the admittedly important and complex constraining and

[176] See Resnick (1994, 1996); Clark (1997).

[177] See, for example, Dennett's (2003) almost sarcastic dismissal of the idea that anyone, including himself, might be a genetic determinist under any useful definition of the term. And yet talk of 'genes for' various characteristics, not to mention innumerable claims about the (genetic) heritability of any number of complex characteristics, including the various 'hardwired' mental modules that are the stock in trade of evolutionary psychology (see, for example, Tooby & Cosmides, 1992), persist in popular and scholarly discussions, despite the apparent denials and qualifications offered by the likes of Dennett.

enabling roles of DNA. Here, for example, is Edelman's summary of the lessons of what he calls topobiology[178]:

> …Cells express genes in time and space to govern morphoregulatory molecules [CAM's, SAM's and CJM's], which in turn control cell movements and cell-to-cell adhesion. These actions take place in groups of cells in proximity, allowing them to exchange further inductive signals. These alter the expression of homeotic genes, which then alter the expression of other genes. *The key players in this topobiological cascade are the cells*, which move, die, divide, release inductive signals or morphogens, link to form new sheets, and repeat variants of the process. Genes control the whole business *indirectly* by governing which morphoregulatory or homeotic product will be expressed. But the actual microscopic fate of a cell is determined by epigenetic events that depend on developmental histories unique to each individual cell in the embryo. (Edelman, 1992, p62; italics added)[179]

Nowhere is the complex and dynamic play of cell movement, differentiation, topological arrangement, division and death in the process of development more evident than in the case of neural development.

In the case of the human brain, the complexity of the developmental task is matched by the sheer size of the task. The adult human brain will eventually consist of up to a hundred billion neurons, and up to ten times as many supporting glial cells. By the time a child is around three years old, somewhere in the region of 10^{14} synapses will have been formed, at rates in the region of 30,000 per second.[180] Even producing an infant brain at birth (far from the finished product) is a Herculean task, made more difficult by the fact that brain development does not proceed in a smooth pattern of successive cell division within an ever-growing ball of progressively differentiating cells. Most neurons need to migrate from where they are formed, by division, to their final position within the developing brain. And yet, for all the complexity and size of the developmental task, the infant brain emerges in a form that has various species-typical characteristics, and ready and able to play various roles in the organism's performance of a range of survival-orientated functions. Brain and neural development thus provide an appropriate case study in which to consider the plausibility of developmental accounts that are genetically deterministic, (primarily) instructionist, and inclined towards assuming/ positing hard-assembled solutions to developmental challenges; as compared to alternative accounts that are more firmly rooted in the dynamics of self-organisation, interaction, selection and soft assembly.

---

[178] The term 'topobiology' is Edelman's chosen term for emphasizing the importance of the *place* or position of a cell within a given sheet of cells within an embryo in influencing the functioning, movement and differentiation of that cell. See especially Edelman's (1988) book *Topobiology*.

[179] 'CAMs' are cell adhesion molecules; 'SAMs' are substrate adhesion molecules; 'CJMs' are cell junctional molecules. See Edelman (1992, pp60-64).

[180] Estimates from Rose (1997), pp144-5.

A rough-and-ready summary of human brain development goes something like this.[181] Following the formation of the neural tube within the embryo, the head end of the tube begins to swell and take the shape of three rough divisions between the fore, mid and hind brain. Precursor cells – the cells that will form the neurons and glial cells – are not formed within the developing brain itself, but must instead detach from the neural tube and migrate relatively vast distances to their eventual destinations. This pattern of migration is repeated by, first, the glial cells, and then the neurons themselves. Finally, similar migratory paths must be taken by neural processes (axons) extending from source/parent cells in one location to target destinations in various other locations in the developing brain. In addition, similar stories unfold for all other parts of the central and peripheral nervous systems, enabling the formation of various sensory and motor pathways between brain and body.

Although many of the details of this story are still the subject of empirical research and debate, it seems that much of the tale will be (as in Edelman's (1992) description of the 'topobiological cascade') a complex story of morphoregulatory molecules (CAM's, SAM's and CJM's), gradients of various trophic factors, and the activities of various homeotic genes (Edelman, 1992; Rose, 1997). This much, at least, appears to be correct in the picture that more gene-centred theorists would like to offer us: genes play a crucial role in the production of these influential molecules, and the manner and timing of this genetic expression is especially important to the degree of *specificity* or uniformity of structure that can be observed in the brains of a given species, including humans. There is, metaphorically speaking, a significant degree of reliably produced *gross* or macro-level 'hardwiring' within the human nervous system, and genes play a significant part in the achievement of this degree of reliable specificity in structure.

Yet we would do well to temper this latter observation on the importance of genetic expression with a number of important qualifications and observations. In particular, we should recall Edelman's (1992) dictum that it is *cells* (and, we might add, cell sheets and other cell arrangements) that are the central players here – to the extent that we desire any candidate to act as 'central player'[182]. Genes play an importantly *indirect* controlling role, in

[181] Based on Rose (1997), pp144-153. See also Edelman (2004), pp28-9.
[182] Again, see Resnick (1994, 1996) and Clark (1997, esp. pp39-45) for criticisms and alternatives to 'centralised thinking'. See also Thelen and Smith (1994, esp. Chapter 1) on the deficiencies of central-cause explanations.

the sense that it is the expression of various genes at different times and in different places that has a crucial enabling and constraining role in development, helping produce a variety of inductive molecules that in turn influence future genetic expression both within an individual cell and in its neighbours. Nothing, however, should make us lose sight of the importance of epigenetic events in determining the fate of each individual cell, as well as collectives of cells; nor, hence, of the importance of (irreversible) individual developmental histories in accounting for both the detailed and overall developmental trajectory taken in the individual brain. In sum, the drama of brain development:

> …is inherently dynamic, plastic, or variable at the level of the fundamental units, the cells. Even in genetically identical twins, the exact same pattern of nerve cells is not found at the same place and time. Yet the collective picture is species-specific because the *overall* constraints acting on the genes are characteristic of that species. (Edelman, 1992, p64)

It is crucial to recognise the extent to which the above-described views of ontogenetic development in general, and neural development in particular, truly stand for a view of development without blueprints[183]. It has become something of a truism to note that the human genome could not possibly encode the specifications for the estimated 10^{14} synapses in the brain; but the gap between acknowledging this truism and properly abandoning the idea that our DNA somehow contains the basic specifications for building a human brain is remarkably wide and resistant to closure. The challenge of coming to grips with the picture of development sketched by the likes of Rose and Edelman is to recognise that there is no single, central orchestrator of the unfolding developmental dance of cells and molecules. Embryonic development, including neural development, is something of a paradigm case for seeing complex interactive dynamics at work, where at any one point in the process, we might shift focus between the importance and influence of such diverse factors as temperature, chemical gradients, the presence or absence of particular molecules, the activity of certain sequences of DNA, the movement of individual cells, or of cell collectives, and topological relations between cells and cell collectives. Depending on our momentary interest and/or focus, one variable might take on the appearance of a 'critical factor' for a while; but then we shift focus and recognise that it is no more critical than a string of others. Now we focus on issues of topological relations as crucial to some unfolding sequence of developmental events; next we note how topology matters in part because of how it relates to and influences temperature and various aspects of a cell's chemical milieu.

[183] To borrow Clark's (1997) phrase.

DNA does not contain the instructions to build an organism, or a brain, any more (or any less) than the topological arrangement of cells in an embryo constitutes a plan or set of instructions for the organ these cells will (quite reliably) form over the course of hours, days and weeks. DNA is critical to development because of its undisputed importance in the formation of proteins. But this is not the kind of critical role that should motivate a 'critical factor' account of development centred on DNA. As has been noted by many a critic of gene-centric thinking, DNA in the absence of a whole lot of functioning cellular machinery is just inert (organic) chemical material. DNA exerts its influence (especially its constraints) on development as part of a multitude of distributed, dynamic and interactive processes that have more in common with the local behaviour of units in self-organising systems than with instructive processes that might build a hard-assembled system via processes of centralised planning and control. Ontogenesis is, from this point of view, much more the self-organising dance of cells than the orchestrated symphony of a genetically-encoded score.

Neural Darwinism

Edelman takes his own version of decentralised, soft-assembly-related developmental processes – most notably, his ideas about selectionist systems – and applies it not only to relatively macro-level accounts of neural development, but to a much more detailed account of neural development that forms, at the same time, the theoretical basis for his account of brain function in general, and consciousness in particular. The significance of such selectionist (rather than instructionist) developmental processes at the level of neural micro-structure is most clearly evident in his 'Neural Darwinism' – more specifically, in his Theory of Neuronal Group Selection (TNGS) that provides the basic underpinnings for all his theorising about consciousness (Edelman, 1987, 1989, 1992, 2004; Edelman & Tononi, 2000)[184].

A crucial element of the anatomical background to Edelman's theory has been anticipated in

[184] It has to be said that Edelman appears to not be well liked in certain philosophical and neuroscientific circles. Horgan (1996), for example, relays a number of disparaging remarks about Edelman made by fellow Nobel Prize winner Francis Crick; and Horgan's account of his own meeting with Edelman is far from flattering. Dennett (1991, 1992, 1995) also enjoys mixing criticisms of Edelman's ideas with comments that can only be regarded as *ad hominem*. While not uncritical of Edelman's work (especially his discussions of neural correlates in Edelman, 2004), my interest here is primarily in the limits to which he attempts to push a selectionist, non-instructionist account of brain development and function.

the discussions above:

> Evidence from developmental studies suggests that the extraordinary anatomical diversity at the finest ramifications of neural networks is an unavoidable consequence of the embryological process. (Edelman, 1992, p82)

In a manner that parallels the characteristics of variant antibody populations within the immune system, Edelman (1992) claims that the embryonic brain is characterised by a massive overabundance of neural connections (in terms of both axonal projections and synapses) that is idiosyncratic (not specified by the genes, not shared by identical twins) and that develops under the influence of multiple micro- and macro-level factors, including irreversible stochastic events such as cell death. Such overabundant and idiosyncratic variation in connectivity patterns provides both (i) a population of variants in the form of multiple possible pathways for neural signals leading to multiple possible 'responses', and (ii) a potentially massive redundancy of pathways, whereby a similar 'response' might issue from multiple pathways. At the same time, as we have already noted, genetic influences and constraints on neural development tend to reliably produce a gross pattern of anatomical organisation that is species specific.

For Edelman, the implications of this anatomical setup, taken together with a number of other considerations, provide strong motivation for adopting a *selectional* account of neural development – his so-called *Neural Darwinism*. Using a digital computer/Turing machine as a prototypical model of a system following instructions, Edelman (1992) contends that (a) the world does not provide a steady stream of unambiguous signals in the manner required by a Turing machine tape; (b) the (neuro) anatomical diversity noted above would constitute variations in hardware wiring that could not be accommodated within a traditional computer system; and (c) an instructional view of brain function would tend to require a signal- and symbol-interpreting homunculus, along with all the problems traditionally associated with homunculi.

In contrast, treating the nervous system as a selectional system "in which matching occurs *ex post facto* on an *already existing* diverse repertoire" (Edelman, 1992, p82, italics in original) exploits the features of anatomical overabundance, diversity and redundancy noted above, in an attempt to explain how the brain comes to categorise the world and guide the adaptive behaviour of an organism in its environment. His own selectional theory, the TNGS (Theory of Neuronal Group Selection), does so by way of three central tenets: developmental

selection, experiential selection, and reentry (Edelman, 1992, 2004; Edelman & Tononi, 2000).

Under *developmental selection*, Edelman has in mind two particular sources of emerging connectivity in the nervous system (see Edelman & Tononi, 2000, p83). On one hand, Edelman points to the kinds of processes (described earlier) through which initial patterns of axonal and synaptic connectivity emerge in a massively overabundant fashion (again, under the constraints of species-specific gross anatomy of the body and nervous system). On the other hand, Edelman sees developmental selection as involving a subsequent initial pruning of these connections based on shared patterns of electrical activity within the neuronal populations. As Edelman and Tononi (2000, p83) put it: "Neurons that fire together, wire together." As a result of relative proximity and gross patterns of shared connectivity within a group of neurons (e.g. neurons tending to receive signals from other parts of the nervous system at the same time), this 'fire together-wire together' logic of correlated firing activity (along with patterns of cell division, cell death, and the growth and withering of axonal processes) will tend to result in neurons within a group being more connected to each other than they are to cells in other groups. Neuronal groups thus form as a significant part of the individual development of the organism, rather than on the basis of some genetic pre-specification or other (potentially centralised) instructional process. Developmental selection is, according to Edelman, especially characteristic of early development, including periods of embryonic development that follow the formation of the gross anatomy of the nervous system; and it gives rise to what Edelman calls the primary repertoire of neuronal groups.

The focus of the second tenet of the TGNS – *experiential selection* – shifts from patterns of anatomical connectivity and selection to the selective strengthening and weakening of *synaptic connections* within the neuronal groups established (and that continue to be shaped) by developmental selection (Edelman & Tononi, 2000). The basis for this selective process is behavioural experience, together with the influence of what Edelman calls the "diffusely projecting value systems" (Edelman & Tononi, 2000, p84) whose activity is constantly subject to change on the basis of successful 'outputs'. As is often the case, Edelman is far from generous with his examples and illustrative explanations, so it will be important to take some time to unpack the different elements and claims at work in experiential selection.

As I interpret Edelman's theory, the distinction between developmental and experiential selection is partly based in anatomy. Embryonic and foetal developmental processes giving rise to the primary repertoire of neuronal groups establish patterns of anatomical links between neurons, including the patterns of available synaptic connections. Amongst these synaptic connections selected under developmental selection, individual connections will be either strengthened or weakened over the course of behavioural experience. As an example, Edelman and Tononi (2000) cite evidence that the boundaries of maps for tactile signals from the fingers can change over time on the basis of varying patterns of finger use. Assuming that Edelman and Tononi (2000) are referring here to functional boundaries within an existing anatomy of interconnected neurons and neuronal groups, the idea is that the strength of synaptic connections (and, thus, of correlated activity) within and between neuronal groups can vary over time based on variations in behavioural 'outputs' (in this case, the use of different fingers) and the value or adaptive success of those 'outputs'. So, superimposed on the developmentally selected *anatomy* of the primary repertoire, we get what Edelman calls the secondary repertoire of connections defined by changes and variations in the *strengths* of populations of synaptic connections.

For purposes of clarity of exposition, I will ignore details about the mechanisms of synaptic strengthening and weakening (as Edelman tends to do in his own expositions of experiential selection – see especially Edelman, 2004 and Edelman & Tononi, 2000). What does require some immediate explication is Edelman's notion of a value system. Edelman (2004) calls these systems of ascending projections (originating in various subcortical nuclei) 'value systems' because of their association with rewards and survival-related responses – that is, they are implicated in signalling value or salience for the organism to diffusely distributed areas of the brain. Each ascending value system is associated with a particular neurotransmitter or neuromodulator capable, when released, of affecting the firing activity of large populations of neurons to which that particular system projects. Bathed in a wash of a given neuromodulator, neurons reached by the axonal projections of a value system will exhibit different probabilities of firing in response to excitatory inputs.[185] The overall effect

[185] While Edelman uses 'neurotransmitter' and 'neuromodulator' more or less interchangeably in relation to the ascending value systems, I prefer the use of the latter term so as to maintain a more consistent distinction between the modulating effects just described, on one hand, and the activities of glutamate, the primary excitatory neurotransmitter of the brain, on the other hand. As Edelman (2004, p25) explains it, the effects of the neuromodulators from the ascending value systems are primarily on "the probability that neurons in the neighbourhood of value-system axons will fire after receiving glutamatergic input." See LeDoux (2002, esp. chapter 3) for a useful overview of neurotransmitters and neuromodulatory chemicals in the brain and body.

of these systems, in the context of Edelman's theory, is thus to "bias neuronal responses affecting both learning and memory and controlling bodily responses necessary for survival" (Edelman, 2004, p25). Some examples of the value systems described by Edelman (2004, p25), and their associated neuromodulators, are summarised in the table below:

Nucleus/Nuclei of Origin	Associated Neuromodulator
Locus Coeruleus	Noradrenaline
Raphé Nucleus	Serotonin
Cholinergic Nuclei	Acetylcholine
Dopaminergic Nuclei	Dopamine

Table 8.1: Ascending Value Systems (based on Edelman, 2004)

The significance of these value systems to a basic understanding of experiential selection is that the strengthening and weakening of synaptic connections associated with behavioural experience takes place under the influence or constraints of the ascending value systems. So while changes in synaptic efficacy might generally be associated with temporal patterns of correlated firing (see, for example, Edelman, 2004, p22), the impact of the value systems will be such as to influence the probabilities of firing patterns over the areas to which a given value system projects, thus acting as a constraint on, or bias in, the processes associated with changes in synaptic efficacy.

Edelman and Tononi (2000) describe the third tenet of the TGNS – *reentry* – as a dynamic process enabling the correlation of selective events across different maps in the brain. Anatomically, reentry is associated with the prevalence of massively parallel and reciprocal pathways between different brain areas (Edelman & Tononi, 2000), most especially within the thalamocortical system (Edelman, 2004). Reentry can act both locally (within a particular map) or globally (between different maps, or even between whole regions) (Edelman & Tononi, 2000). In the context of the TGNS, and in Edelman's overall account of consciousness, reentry is a process of which much explanatory work is expected, as

evidenced in the following extended quote[186]:

> Reentry allows an animal with a variable and uniquely individual nervous system to partition an unlabeled world into objects and events in the absence of a homunculus or computer program… [Reentry] leads to the synchronization of the activity of neuronal groups in different brain maps, binding them into circuits capable of temporally coherent output. Reentry is thus the central mechanism by which the spatiotemporal coordination of diverse sensory and motor events takes place. (Edelman & Tononi, 2000, p85)

In short, Edelman sees reentry as promising both a solution to the binding problem, as well as a non-instructive, non-feedback-based mechanism through which categorisation (across different neural maps and sensory modalities) can be achieved in the absence of central executive control *a la* a homunculus, Turing tape reader or computer CPU.

The complexities of Edelman's account of reentry, not to mention the inaccessibility of much of his prose, argue against a more detailed unpacking of the concept at this point. In the context of the current chapter, however, it is important to unpack Edelman's specific claims about the differences between reentry and feedback. Edelman and Tononi (2000) draw the contrast in this way:

> Feedback occurs along a *single* fixed loop made of reciprocal connections using previous *instructionally* derived information for control and correction, such as an error signal. In contrast, reentry occurs in selectional systems across *multiple* parallel paths where information is not prespecified. (Edelman & Tononi, 2000, p85, italics in original)

I take it that Edelman wants to highlight the following aspects of the contrast between feedback and reentry. With feedback, whether in a bodily homeostatic system or in a machine, there is typically a single path or loop involving the transfer of information to which parts of the system are instructively sensitive. So the cells in a hormone-producing gland are differentiated or specialised in such a way that they can be described as instructively sensitive to concentrations of that hormone (or some other molecules related to the relevant hormone system), producing more of the hormone or less of the hormone according to the changes in concentration of the relevant molecules detected by the gland. An initial response to a given change in concentration (say, an increase in hormone production/release) is followed further down the line of the feedback loop by a correlated change in detected concentration that, in turn, corrects the system by reducing hormone production/release. The system has an instructive control function built into it through its capacity for differential response to a particular kind of chemical information.

[186] Searle's (1997) discussion of Edelman's work is titled 'Gerald Edelman and Reentry Mapping', highlighting the significance Searle attaches to reentry within his interpretation of Edelman's theory.

In contrast, Edelman wants to emphasise that reentry involves multiple parallel pathways of concurrent neural activity (hence not a single loop) that has the ability to synchronise firing activity (locally and more globally) in neural maps without the reentrant signals having a specific instructive or information-transfer function.

It is not difficult to see why critics might want to nevertheless label this reentrant signalling as feedback: on the surface, an interactive signalling process, leading to coordinated activity in different systems or parts of a system, just sounds like it must be a kind of feedback mechanism. To use a Dennett-like turn of phrase, if it does what feedback does, it must be feedback. But I think this response refuses to take seriously Edelman's larger project of articulating a global brain theory that tries to assume as little as possible about the need for *information transfer* and *instructive control* as a basis for mind-brain functioning, despite the temptation to think that the brain *qua* biological information processor should be open to description and explanation in precisely these instructionist terms.

Specifically, I think Edelman is pushing us to resist *one particular way* of thinking about reentrant signalling, in favour of his more radically selectionist interpretation. For example, Edelman and Tononi (2000) assert that reentry can enable a synthesis of neural responses in different submodalities, such as colour and motion in the visual system. It is, I think, clear that it is all too easy to reinterpret this idea in terms of neurons in one part of the brain *telling* neurons in another part of the brain what they are up to, where 'telling' is to be unpacked on a model of information transfer – the colour information processed and decoded in the colour processing areas is being communicated and exchanged with the motion information decoded by the motion processing areas. Instructionist thinking of this kind will then all too easily lead to questions about later delivery (or integration) of this information further down the line – perhaps in or to an association area, or the frontal lobes – with the attendant risk of then introducing a homuncular central executive 'located' in some part of the brain.

Whatever the temptations – even the apparent utility – of speaking in terms of information transfer of this kind, I take it that Edelman thinks of reentry as specifically not requiring information-based instruction. In somewhat simplified terms, if reentrant signalling *signals* anything, it signals a much more basic 'message' – something more like 'there is activity in a neuronal group or map at the other end of this reentrant channel.' Moreover, such 'signalling' between two brain areas is concurrent and bidirectional. Edelman's theory (and his computer

simulations of reentry in different generations of his Darwin robots) suggest that this kind of non-instructive signalling, under constraints of value (and alongside mechanisms like the fire-together-wire-together logic of developmental selection, together with the capacity for strengthening and weakening of synaptic connections characteristic of experiential selection) can yield synchronised and coherent categorising activity across (real or simulated) neural maps. Crucially, Edelman sees categorisation as a *global* activity that emerges from these neural and bodily dynamics – and I take it that he means that, for the most part, information *emerges at this global level*. Reentry, then, does not involve the instructional information transfer typically characteristic of feedback because, at the level of function described by the TGNS, reentry occurs at a *sub-informational* level. Hence Edelman and Tononi's (2000, p85) claim, quoted above, that reentry is critical to understanding how organisms like ourselves "with a variable and uniquely individual nervous system [can] partition an unlabeled world into objects and events in the absence of a homunculus or computer program."

Salient examples of what reentry can help enable in the absence of a programmer or specific instructive or learning algorithm come from various computer-simulated systems that Edelman and his colleagues have constructed and tested over a number of years. Reeke and Edelman (1984), for example, describe a computer simulation (aptly named Darwin II) constructed on the basis of the TNGS whose task was to learn categories (specifically letters of the alphabet) based on exposure to variant tokens of these categories, and without specific instruction or programming. As Thelen and Smith (1994) highlight in their discussion of this particular simulation, the learning of letter categories is generally assumed to be a clear instance of explicit teaching – children will learn, through explicit instruction and correction, which marks on a page are instances of a particular letter and which ones are not. Reeke and Edelman's (1984) challenge was thus to produce an artefact whose performance in learning to discriminate letters of alphabet without such specific instruction and error-based feedback, using the principles of the TNGS, could act as a demonstration proof for the possibility such learning.

I will not delve too deeply into the details of the simulation. In essence, Reeke and Edelman (1984) constructed two parallel networks linked to the same 'optical' input array – a feature detection network (topographically-sensitive to lines of different orientations, curves, etc.) and a network involving a tracing mechanism (analogous to the eye scanning an array and tracing the outlines of objects encountered in the visual scene). Each network is, in one sense,

independent since each has its own method of sampling and registering whatever is present in the input array. Moreover, the characteristic sensitivities of each network mean that, on their own, they have different response tendencies. The topographical feature analysis of the feature detection network makes it likely to produce unique responses[187] to each stimulus pattern in the input array; whereas the sensitivity of the tracing network to the mere presence of lines and line junctions disposes it to producing the same response to a class of inputs irrespective of various transformations and/or distortions. On the other hand, in line with the TNGS, the system as a whole is designed to allow reentry both within and between these two networks.[188]

Exactly what claims can be made on the basis of the results reported by Reeke and Edelman (1984) is a little unclear. Reeke and Edelman (1984, p198) are of the opinion that the various experiments they report "demonstrate that a network based on a selective principle [those of the TNGS] can function in the absence of forced learning or an *a priori* program to give recognition, classification, generalization, and association." Of course, what is meant in each case by 'recognition', etc. is an analogue of recognition represented by patterns and changes in activity in the system that Reeke and Edelman interpret as recognition. Nevertheless, these are remarkable achievements for a system that has no explicit instructive function for learning letters, and that is not employing any evaluation of outputs from the system.

Thelen and Smith (1994, p169) are even more enthusiastic about the outcomes of the study:[189]

> Reeke and Edelman showed that this device could learn and generalize letter categories. The device *teaches itself* to recognize letters without making any externally evaluated responses. No one needs to tell it that all As are As for it to discover the similarities that exist between As. The intelligence of the device is in the simultaneous self-organizing activity of the three maps[190]; the intelligence is in the pattern of activity of the whole.

---

[187] Where 'response' means something like 'pattern of activity within the network' rather than a specific output of some kind. It is important to note that Darwin II does not involve any externally evaluated outputs.

[188] The effects of simulated reentry being especially evident when comparing the system's functioning both with and without internetwork reentry – i.e. activity in the networks can be compared when they are functioning independently and when they are functioning with reentry.

[189] One cannot help but suspect, given various details contained in their exposition not evident in the original published study, and the general accessibility of their presentation, that Thelen and Smith (1994) drew on more than just the published article by Reeke and Edelman (1984) – perhaps personal communication with Edelman, whose theory forms an important part of Thelen and Smith's (1994) own project. I have, however, been cautious in my use of Thelen and Smith's (1994) presentation for fear of uncritically adopting their enthusiastic interpretations.

[190] Thelen and Smith (1994) are referring to the feature detection mapping, the trace mapping, and the reentrant mapping occurring between these networks.

Even if we treat this as an optimistic extrapolation of what Reeke and Edelman were in fact able to demonstrate using Darwin II, the possibility of such a demonstration proof of the self-organising power of selectionist networks that incorporate reentry is itself significant. The exciting idea is that a system could learn to categorise, without being taught to categorise, as an emergent feature of the self-organising dynamics of the system. The activities of two or more sub-systems that each represent independent samplings of a stimulus input are correlated in real time with the result that the sub-systems 'educate' each other. Of course, the system (like a human brain) embodies all kinds of features and properties – it has its own intrinsic dynamics – and these are important to the system's capacity to learn and categorise. But learning emerges from what happens when a system with such intrinsic dynamics is simply exposed to instances of As and Bs. An external programmer or teacher, or internal controller or instructive (e.g. error correcting) algorithm, is unnecessary.[191]

Displacement

A final example of a developmental process that illustrates ideas of soft assembly, decentralised control, and development without blueprints, again specifically in the context of brain development and function, is Terence Deacon's (1997) notion of displacement. Like Edelman, Deacon believes that many of the more important processes of brain development need to be understood in more or less Darwinian terms[192], specifically in terms of processes of competition and selection amongst available alternatives. Deacon's account of displacement is a relatively simple application of these ideas to an account of the (ongoing) development of connecting pathways amongst diffuse areas in the brain.

As outlined earlier, patterns of cell migration and patterns of axonal projection are critical aspects of brain development. Deacon's (1997) focus in talking about displacement as a critical selective process in brain development is on the latter. For neurons sending out projections to other neurons, whether locally or not, the destination of their projecting axons

[191] Edelman (1989) provides an extended and complex discussion of simulations based on his Reentrant Cortical Integration (RCI) model for certain aspects of early visual processing. Detailed discussion of this example would occupy too much space here, given the 'take-home' message already extracted. In short, Edelman (1989) claims that reentrant signaling between segregated circuits (analogues of areas devoted mainly to detecting orientation, occlusion and motion) could, for example, resolve conflicts in the detection of illusory contours, or combine motion cues (direction and direction discontinuity) and occlusion cues to synthesise illusory contours in a combined illusion of perceived shape.

[192] Deacon (1997, pp473-4) sees his view as having much in common with Edelman's Neural Darwinism. He contrasts his and Edelman's positions, however, in terms of (i) assumptions about the sources of variation available for selection, and (ii) Deacon's special emphasis global biases that emerge as a result of relatively large-scale quantitative relationships between neural regions. See the discussion that follows in the text.

is, of course, the dendritic processes of some other brain area/s or individual neurons. These dendritic processes are, in turn, often the site for many hundreds or thousands of axonal terminations; and, for any particular projecting axon, there is competition for synaptic space on these dendrites with axons stemming from local neurons, other local and diffuse brain areas, as well as axons that project from the same area as the axon itself. In the course of development, the overabundance of connections at a particular dendritic site will be pruned through processes such as cell death, as well as axonal withering of projections whose synapses are not strengthened through (correlated) use[193].

Displacement (in Deacon's sense) is a straightforward quantitative effect of a competitive/ selective mechanism of axonal pathway development, given suitable variations in the *relative sizes* of projecting brain areas 'targetting' any single dendritic destination. In purely statistical terms, it is clear that a larger projecting area will send more axonal projections to a given dendritic site than a smaller projecting area with fewer axons. Thus, on average, projections from the larger area will tend to 'crowd out' projections from the smaller area, with more of the larger area's projections surviving into later development and adulthood through sheer weight of selective forces favouring the larger numbers of potentially 'successful' projections. Under such circumstances, a relatively larger area of axonal origination is said to *displace* connections that might otherwise have been formed by competing axons from either local neurons or other projecting brain regions.

In principle, the significance of displacement as a developmental process is that it provides a mechanism that could account for quite radical changes in connectivity patterns within the brain *without* requiring any specific genetic mutation that 'codes for' this novelty or change (together with any associated changes in functionality). That is, the potential developmental effects of displacement suggest that genetic mutations associated with, for example, a change in the *relative size* of one brain area, and/or the *relative pace* of development of a brain area, might in themselves lead to novel patterns of connectivity (and function) that cannot be meaningfully attributed to the mutation itself – at least, not in the sense that the mutation could be said to *code for* the change in connectivity[194]. Deacon's (1997) favoured example, involving the origins of human linguistic capacity, is the increasing encephalisation of the

[193] That is, primarily by way of processes described by Edelman as *developmental selection*.
[194] To the extent that we might ever want to talk about a gene, or a genetic mutation, 'coding' for anything in particular other than, perhaps, a protein of some kind.

primate and hominid lines that, through displacement, leads to increasing forms and degrees of connectivity between the cortex and various subcortical structures.

Brain disproportion, displacement and cortical control

Deacon (1997) argues (and provides empirical evidence) that the increase in relative size of the tertiary and frontal areas of the cortex in hominid and human brains is, by way of processes like displacement, significantly responsible for the increasing degree of cerebral connectivity to, and resulting cortical influence over, subcortical structures that control breathing, and the muscles of the larynx and tongue (these examples being, of course, vital to the emergence of human speech).

Deacon summarises his view of this trend in mammalian and primate evolution as follows:

> As a result of the reduction in proportions of the postcranial body compared to the head and brain there has been a change in the relative proportions between the forebrain as compared to the brain stem and spinal cord, and this embryological shift in neural proportions is a recipe for displacement. With so many more descending axons vying for space in the primate motor system, the more numerous cortical axons displace the less numerous local connections, since these displaced connections arose from systems that were scaled for a smaller body. In non-primate brains, the initially overexuberant and somewhat non-specific cortical projections to these brain stem motor nuclei are outcompeted by local projections and pruned back during development to leave only those projecting to premotor regions of the brain stem and spinal cord. In primate brains, the initial cortical projections are so numerous that they outcompete the local connections and persist in far greater numbers in many additional motor nuclei. (Deacon, 1997, pp248-9)

In the case of human brains, Deacon observes that an additional increase in cortical/ brain stem disproportion over our primate relatives and ancestors would likely lead to an even greater invasion and recruitment of subcortical motor nuclei, and consequent increases in cortical influence and control:

> [Cortical axons] will almost certainly increase in proportions in face and tongue muscle nuclei, with the consequence that the voluntary control of these systems will be greater than in other primates. In addition, however, the more extensive human cortical projections have probably also invaded nuclei in the brain stem and neurons in the spinal cord that even primates do not have voluntary control over: nuclei controlling visceral muscle systems. (Deacon, 1997, p249)

Of these latter nuclei/ neurons, those that control the larynx and breathing are of particular interest to Deacon because of their relevance to speech – something that requires a degree of control over vocalisation not evident in our primate cousins.

Consider, then, relationship between breathing, swallowing and laryngeal control. As long as we are not vocalising, breathing is mostly an automatic behaviour in humans, as is most clearly evident during periods of sleep and unconsciousness. Similarly, control of the larynx

is also largely automatic, as evidenced in the (usually) reliable swallowing reflex that safely shuts off the air passages to allow fluids or solids to pass into the oesophagus. Breathing and the movements of the larynx in swallowing are, then, largely automatic, coordinated functions that carry on beyond the reach of (or need for) cortical control. However, humans also display a remarkable degree of control over breathing and laryngeal movement, at least when vocalising. Deacon's (1997) claim is that such potential for control is, in significant part, a result of the 'invasion' of cortical projections into nuclei that, in other species, would tend towards more exclusively autonomous functioning under an array of subcortical influences. Deacon describes this pattern of increasing cortical influence, via mechanisms like displacement, as a 'leveraged takeover' of motor control via increased cortical projections.

We will revisit this idea of a leveraged takeover, and the novel opportunities it presents for the influence and control of action, in discussing Merlin Donald's (2001) work in Chapter 10. For present purposes, it is the mechanism and process of displacement itself that is of interest as a further example of decentralised control and soft assembly. Specifically, the mechanism of displacement offers a strong basis for challenging various ideas about the potential controlling power of genes in the (hard) assembly of brains that exhibit distinctive species-specific patterns of connectivity.

Let us suppose that Deacon (1997) is correct in his claims about the dramatic increase in proportions of cortical-subcortical axonal projections in human brains, as compared to our primate cousins. Let us further suppose that these distinctive patterns of connectivity are so reliably reproduced in development that they have a strong and robust statistical association with a particular genetic feature not found in our primate cousins. From a statistical point of view, then, the distinctive patterns of cortical-subcortical connectivity might appear to be 'attributable' to the distinctive genetic characteristic. A critical question, at this juncture, is this: what exactly should we attribute to this genetic factor? How should we understand the genetic 'influence' associated with our distinctive neuroanatomy?

The answer that would be encouraged by our tendency towards thinking in terms of centralised control, and associated processes of hard assembly, would be one in which the relevant genetic feature is thought of as *coding for* the relevant neuroanatomical pattern. That is, the genetic feature in some way encodes *instructions* for 'wiring' a brain in such a way as

to yield the requisite pattern of increased cortical-subcortical projections. The ontogenetic emergence of the pattern of projections would thus be thought of as under the (centralised) control of a set of genetic instructions. The locus of control is not spatially centralised, seeing as all cells contain the relevant genetic instructions. But it is centralised in the functional sense of being under the influence of a single critical factor that is identical in all of the sets of genetic instructions, and that therefore directs the emerging 'wiring' activity, as it were, with a common (instructive) voice.

Displacement provides us with an empirically and conceptually plausible alternative to this interpretation of how brains and bodies are assembled. Given the statistical regularities involved, Deacon (1997) has shown how a distinctive pattern of neural pathways can reliably emerge on the basis of changes in the relative sizes, proportions and rates of development of different brain regions. The pattern of *connectivity* does not need to be coded or planned anywhere, or instructively controlled by some functionally 'central' critical factor. Instead, distinctive species-wide macro-anatomy and functionality can emerge from the interplay of a number of developmental factors ('equal partners', to use Thelen and Smith's, 1994, and Clark's, 1997, phrase), including the statistical logic of a selection-based mechanism such as displacement. There are, on this view, mechanisms of control in development that produce order and coherence in a reliable and coordinated fashion; but that control may well (like processes of displacement) be distributed, decentralised, and involve a number of interacting factors.

The Significance of Self-Organisation, Soft Assembly and Selectionist Systems

We have taken an extended excursion into a variety of domains in which this collective of related ideas – self-organisation, decentralised and distributed control, soft assembly, selectionism – can be encountered in one or other form. This is not, however, merely some loose collection of 'sexy' ideas or faddish concepts that happen to be doing the rounds in parts of the biological and cognitive sciences. Instead, they represent an interconnected group of concepts that offer considerable explanatory power in various domains of application; and the range of examples described in the current chapter serves as testament to the potential range of applications for these ideas, especially within the biological sciences.

Moreover, some of the systems to which these explanatory concepts have been applied are not just any biological systems. They have notably been applied to systems that are found in

human agents. These ideas have potential relevance for how we understand genetic processes and genetic expression in general, for how we understand the adaptive functioning of our immune systems, for how we understand and explain the ontogenetic processes involved in overall brain development, as well as for certain detailed accounts of neuroanatomical development and associated functioning. And, in Edelman's hands, they form an integral part of his attempt to explain primary consciousness.

A first conclusion to draw from this extended discussion is that, at the very least, we can be sure that there is more than one variety or form of control at work in a complex biological system such as a human organism. While each of the applications and theories we have discussed is empirically defeasible, it is an empirically reasonable bet that one or more of them will turn out to be largely correct explanatory accounts of the systems that are their targets of interest. And while this modest but important conclusion cannot, in isolation, settle any of the debates about conscious agency and control raised in Chapter 6, it at least motivates caution, and the need for careful empirical inquiry, when it is claimed or assumed that conscious (executive) control fits the kind of centralised, hard-assembled model that is at work in the sceptical arguments of Chapter 6.

Before attempting to develop this conclusion further, I think we can anticipate at least two objections. On the first objection, a critic might protest that they cannot see the relevance of all these concepts and examples to the issue of executive control. Given the conclusion just offered, the discussion does not stand as an argument against the possibility that human agency involves the superimposition of a genuinely centralised form of executive, instructive control on top of whatever other (subpersonal) systems of control we might find within our organisms, whether these latter be decentralised, soft-assembled, self-organising, or not. And if it were to turn out that human agency, or our assumptions about a distinctive form of human agency, requires a more centralised form of executive control, then it is not clear how pointing to the existence of other varieties of control in biological (and other) systems could be of assistance in defending against the sceptical conclusions of Chapter 6.

The second objection is of a different kind. On this objection, a critic might wonder whether the ideas and examples discussed in the current chapter are not *extra* ammunition for revising our assumptions about conscious agency and control along precisely the lines of the sceptical positions sketched in Chapter 6. The critic might argue that ideas about self-organisation and

decentralised control just make the 'revisionist' views of agency proposed by the likes of Bargh and Wegner all the more plausible, precisely because they highlight ways in which systems can exhibit coordinated and adaptive responses without these needing to be initiated, orchestrated and managed by some well-informed centralised (conscious) controller. Framing the behaviour of such systems as self-organising behaviour, soft-assembled solutions, decentralised control mechanisms, or in terms of unconscious and sub-personal processes, makes little difference to the conclusion, which is that we are mistaken in thinking that we exert a significant degree of conscious control over our mental and behavioural existence.

Another way of framing this second objection would be to claim that an overly strong emphasis on decentralised and distributed systems of control, together with an emphasis on the power of self-organising systems made up of essentially 'dumb' units responding to local conditions on basis of relatively simple rules, all sounds very much like Daniel Dennett's more sceptical views on consciousness and the self. After all, what Dennett seems to offer with one hand (e.g. the more constructive, agency-friendly aspects of works like *Elbow Room* and *Freedom Evolves*) he often takes away with the other hand (especially when in sceptical mode about consciousness, as in *Consciousness Explained*; but even when banging on about Cartesian thinking and the legacy of the Cartesian Theatre in works like *Freedom Evolves*). In short, when Dennett (2003) indicates that he is happy to smear the self of agency around in time and space, he seems a little too happy to allow it to fragment and disappear amongst a host of non-conscious, sub-personal processes, leaving the rest of us (who attach some value to a less sceptical view of consciousness) feeling deeply uneasy. The objection, then, is that the discussions of this chapter so far seem to point towards a Dennett-like *dismantling of the agent*, rather than towards a suitably nuanced and complex understanding of agency that can nevertheless sustain claims of agent unity, integration, and the importance and efficacy of consciousness.

Both these objections mistake the immediate point of the preceding discussion, as well as of the larger project that is being advanced. Recall that the central conclusion of this chapter, so far, is that we can be sure that there is more than one variety or form of control at work in complex biological systems such as a human organism. As against our first critic, this suggests that it would be a conceptually and theoretically risky strategy to put forward only one model of distinctively human executive control that is strongly instructive and centralised, when we have good reasons for thinking that any such executive control system

must *at the very least* be compatible and capable of interfacing with varieties of systems that do not fit such a highly centralised model.

Consider again, for example, Thelen and Smith's (1994) accounts of the development of reaching behaviour and of walking. Since these are, as noted, empirically-defeasible accounts, let us assume that they turn out to be largely correct explanations of these achievements in human children, meaning that the solutions to these developmental challenges are soft-assembled strategies that employ decentralised and distributed forms of control in service of overarching intentional goals. Our first critic may well imagine superimposing a system of executive control over these important but modest successes in achieving intentional bodily control, and they mistake the direction of my project if they imagine that I do not want to say more about distinctively human agency, including appropriate notions of executive control that can respect empirical facts about our natures, abilities and limitations, while also serving as a basis to secure (some of) our claims to being free agents. What this critic should recognise is that it would be explanatorily peculiar to imagine that any superimposed executive control system would somehow have to re-solve the problems of coordinating reaching and walking behaviour by replacing decentralised and soft-assembled solutions with strongly centralised, hard-assembled, instructive solutions[195].

So, again, the lesson of the current chapter is that we should not expect conscious influence and control over action to just neatly fit one model or image of how such control is achieved – not just because we want to hedge our empirical bets, but because there is good reason to think that there are many different systems and levels of control at work within our bodies. And this matters greatly to the larger project of countering the sceptical arguments of Chapter 6, because the self-declared targets of sceptics like Bargh and Wegner, and the 'naïve' model

[195] For example, the ideas about kinematic imagination (Donald, 2001) discussed in Chapter 10 are properly associated with an increase in executive control. Moreover, in drawing on Deacon's (1997) ideas about displacement and the 'leveraged takeover' through which various subcortical systems come increasingly under the influence of cortical processing in the human brain, the employment of kinematic imagination as a means for developing a skilled performance *does* involve superimposing novel forms of influence, modulation and control over what were previously largely autonomous neural systems of bodily control. The critical issue, in the light of the ideas and arguments of this chapter, is one of emphasis. Increases in cortical influence and modulation over subcortical control structures *form part of* a novel form of control system. But the control system thus realised is not a cortical control system, but rather a system that *includes* cortical and subcortical structures and processes that, in dynamic interaction with each other, make novel forms of embodied performance possible. (To treat the new layer of cortical influence as *the* critical control system is to once again succumb to the temptations of centralised thinking (Resnick, 1994, 1996).) On this view, a process like soft assembly remains amongst the critical mechanisms for assembling and perfecting skilled performance in pursuit of intentional goals across the lifespan of the individual – it is not a process that is or needs to be supplanted by more distinctively adult human capacities for learning.

of agency purportedly under investigation in Libet's studies, all depend on the idea that conscious control *fits one model* of how intentional behaviour needs to be caused if we are to claim effective conscious control over ourselves and our activities as agents.

This same lesson, however, also points to something wrong with the objection made by the second critic. The second critic is, in effect, suggesting that mechanisms of self-organising behaviour are so powerful and versatile that once we recognise this, we will find increasingly little motivation to posit *any* other mechanisms of control, even in complex systems like human agents. But this suggestion *also* would have us believe that we are only going to encounter one basic form of system organisation and control in certain complex biological systems and, more specifically, that the form of self-organising behaviour we will find will be of a kind with, say, ants and their collective behaviour – i.e. it will involve essentially 'dumb' units blindly responding to local conditions according to simple 'rules', such that we can talk about the intentions of the system and the knowledge distributed in the system, but all of this would be metaphorical psychology at best. The appropriate response to this one-track (reductive, dissipative) view of organisation and control is, for a start, to point out that it would be surprising in the extreme to find only one type of control at work in a complex system like an embodied human agent. To suppose this is no more reasonable than supposing that all conscious intentional behaviour involves one kind of special executive control mechanism.

Yet our second critic requires a stronger response than just the above corrective. The second critic is, after all, expressing an eagerness to concede that the important concepts and examples discussed earlier are persuasive with regard to the relevance of these kinds of mechanisms to human functioning. What they are asking for is a good reason to *stop* pushing this kind of explanatory project to the point where *all* organisation, coordination and control of behaviour in human agents is achieved by such decentralised and distributed means. In other words, how do we avoid the process of dismantling the conscious agent once we have recognised the power of self-organising systems and decentralised systems of control?

There is a real tension here between, on the one hand, the powerful dynamics of self-organisation that promise to explain the behaviour of a range of systems without needing to resort to any kind of central planner, coordinator, controller or knower and, on the other hand, the desire (articulated in Chapter 5) to avoid the sidelining of consciousness and conscious

awareness in ways that might suggest we are, in the end, much more like AAs than suitably self-aware free agents who exert meaningful degrees of control over our lives. It is a tension that has been noted by others who have been intrigued by, and persuaded of, the potential importance of self-organisation to understanding human agency, while remaining equally convinced (unlike our imagined sceptical critic) that there is an important place for meaningful, indeed ineliminable, talk about self-governance too[196]. How should we balance these perspectives in such a way that an argument for self-governance does not run afoul of precisely the mistakes I have been cautioning against earlier in this chapter?

This brings to a head one of the deepest divisions regarding the appropriate characterisation and explanation of human behaviour – namely whether or not we are willing and able to give or reserve a distinctive role to/ for consciousness when it comes to a complete and adequate account of human agency. At one extreme, we have people like Schroeter (2004), and probably many agent-causal libertarians, who want to identify the agent that wields basic executive control with the conscious self. At the other extreme, we have the likes of Dennett and his followers[197], probably along with the likes of Bargh, who are quite happy to see standard notions of consciousness and conscious selfhood buried along with Cartesian dualism, the Cartesian Theatre, and any number of other supposedly 'pre-scientific' ideas we have had about the mind and agency[198]. In between, there are those (like me) who want to articulate exactly how and why consciousness matters so much to our agency while recognising that (a) consciousness cannot license 'ghost in the machine' explanations, but also (b) explanations involving consciousness may in the end look very different to explanations to be found everywhere else in science, precisely because consciousness is so distinctive; and because it matters.

From this perspective, our second critic can be accused of playing the 'more-of-the-same' card that is familiar to readers of Dennettian and other accounts of consciousness, in which it is alleged that something that we thought distinctive of consciousness and conscious agents

[196] I am thinking in particular of the work of Jennan Ismael. See her *The Situated Self* (2007).
[197] See, for example, various enthusiastic and/or sympathetic commentaries on Dennett's work in Brook and Ross (2002). Arguably the most enthusiastic follower of Dennett is Susan Blackmore. See, for example, Blackmore (1999) as well as her contributions to the interviews she conducted in Blackmore (2005).
[198] It is far more difficult to fit someone like Wegner into this picture for, while Wegner pushes the idea that our experience of conscious will is always illusory, his account depends on acknowledging that our brains go to the expense of generating this distinctive conscious experience *for us* and, moreover, that there may be benefits to our having this experience.

like ourselves is really just some complex, perhaps iterated combination of something else that is not at all distinctive of consciousness and conscious agents – a kind of "We-are-just-really-sophisticated-thermostats" view of intentionality and consciousness. In this case, the suggestion is that the 'real' work of agency we might have wanted to associate with consciousness is really *just* being done by a complex combination of interacting self-organising systems. So the control we find in conscious agents like ourselves *really* turns out to just be *more of the same* kind of control we can find in ant colonies, flocks of birds, and slime molds.

This strain of resistance to any special role for consciousness is well illustrated by a to-and-fro exchange between Dennett and Gallagher[199] regarding the appropriate interpretation of and response to Libet's data on the timing of conscious intentions. In a postscript (on free will) to a 1998 interview with Michael Gazzaniga in the *Journal of Consciousness Studies*, Gallagher had written:

> I think that this problem can be solved as long as we do not think of free will as a momentary act. Once we understand that deliberation and decision are processes that are spread out over time, even, in some cases, very short amounts of time, then there is plenty of room for conscious components that are more than accessories after the fact. (Gazzaniga & Gallagher, 1998, p.715)

The problem Gallagher is referring to here is the sceptical interpretation of Libet's data on which it seems, *a la* Wegner, that the brain both takes a decision before consciousness has any chance to get in on the act, *and* then further plays a trick of making us think that consciousness decided the matter afterall. Dennett (2003) thinks that Gallagher is mostly on the right track, both in opposing the idea that free will somehow consists in momentary acts, and in wanting to spread out the agency involved in deliberation and choice over time. But unsurprisingly, Dennett latches on to the prominence Gallagher wants to give to conscious components and processes as just one more hangover of Cartesianism. He chastises Gallagher for supposedly going on to claim that "if the feedback is all unconscious, it will be 'deterministic' but if it is conscious, it won't be" (Dennett, 2003, p242 n. 3). As if to remind us how easy it is to fall into the traps he has so carefully identified for us, Dennett ends by commenting that "Cartesian thinking dies hard" (ibid.).

[199] The initial 'exchange' involved Dennett (2003) commenting on Gazzaniga and Gallagher (1998) in a footnote. Gallagher (2005) quotes this footnote, and then responds to Dennett's commentary. I encountered the different parts of the exchange, in essence independently, in both Dennett (2003) and then in Gallagher (2005).

To be fair, Gallagher's reference to "deterministic" loops in the production of action was a poor choice of phrasing. What he said was this:

> There is some feedback that is irrelevant to the issue of free will. If components involved in a feedback process are limited to completely physical events such as nonconscious brain events, the loop is completely deterministic. (Gazzaniga & Gallagher, 1998, p.715)

If I understand Gallagher correctly, what he meant was that if the feedback loops associated with behaviour are entirely non-conscious events, and particularly events that might be given an exclusively 'physical' description, then the effects of that feedback loop on unfolding behaviour would be entirely determined by the laws and/or regularities governing those types of events. The intended contrast was not supposed to be between the determinism of non-conscious brain (and other physical) events and the indeterminism of conscious events. Rather, the contrast was supposed to point to the importance of conscious feedback loops *qua* conscious processes in the production of intentional behaviour that is the proper subject of questions about free will.

It is worth emphasising just how close Gallagher and Dennett are to each other as regards their basic response to Libet's data. In *Freedom Evolves*, Dennett concludes his discussion of Libet's studies thus:

> When we remove the Cartesian bottleneck, and with it the commitment to the ideal of the mythic time *t*, the instant when conscious decision happens, Libet's discovery of a 100-millisecond veto window evaporates. We can see that our free will, like all our other mental powers, has to be smeared out over time, not measured in instants. (Dennett, 2003, pp.241-2)

In 1998, Gallagher had recorded his own similar take on the appropriate conclusion to draw about Libet's work: "What we call free will cannot be conceived as something instantaneous, a knife-edge moment located between being undecided and being decided" (Gazzaniga & Gallagher, 1998, p.717).

Dennett (1991, 2003) is, however, committed to his sceptical 'anti-Cartesian' view that there is no 'finishing line' for things arriving or entering into consciousness. Or, as he might put it using his Cartesian Theatre metaphor, there is no moment when things take the stage in the Cartesian Theatre (for the benefit of the homuncular audience seated in the theatre). Because he takes this sceptical view, Dennett does not and cannot make a principled distinction between conscious processes and un- or non-conscious processes. Exactly what marks something as 'conscious' on Dennett's view is not always clear, but in *Consciousness Explained* (1991) it comes to something like 'featuring in a report in response to a question or

probe.' And we can see this sceptical stance clearly at work in the latter part of his response to Libet in *Freedom Evolves*:

> Once you can see yourself from [the perspective in which 'you' are not some extensionless point but are instead distributed in both space and time in the brain], you can dismiss the heretofore compelling concept of a mental activity that is *un*consciously *begun* and then only later "enters consciousness" (where *you* are eagerly waiting to access it). This is an illusion since many of the reactions *you* have to that mental activity are initiated at the earlier time… (Dennett, 2003, p.242, italics in original)

For me, Gallagher, and everyone else who doesn't share Dennett's sceptical stance on consciousness and the self, there is all the difference in the world between something that has crossed the threshold into consciousness and things that haven't[200], whatever the puzzles and problems that this idea might have associated with it. Dennett's view is effectively a 'more-of-the-same' position in which distinctively conscious processes (such as Gallagher's conscious feedback loops) disappear from view to become merely one amongst the many parallel processes going on at any one time in the brain of the agent. The 'conscious' processes might be ones that get reported on in response to a probe, or that feature as a narrative strand in an account spun by the self-as-centre-of-narrative-gravity[201]; but that is *all* that marks them out as conscious (in contrast to any of the other processes running alongside them). This does not explain consciousness and conscious processes – it simply eliminates these as phenomena that require scientific (and philosophical) description and explanation[202].

If we reject, as we should, this kind of more-of-the-same reductionism (or eliminativism) that refuses to mark out distinctively conscious processes, and we do so both generally and specifically in relation to the central issue of this chapter, namely *control*, then we can see why the central lesson outlined earlier *does indeed argue against replacing distinctively conscious control and agency with a mere concatenation of self-organising systems and sub-systems*. That central lesson implies not only that we should *not* look for conscious control to be in the business of re-doing what is already being done by way of more decentralised and distributed means (my response to the first critic), but also that we should now have a clearer view of where distinctively conscious processes of control in agency could and should fit in,

[200] See, for example, Baars (1994, 1997), on just some of the many differences between conscious and non-conscious processes.
[201] On the notion of the self as centre of narrative gravity, see Dennett (1991), especially Chapters 13 and 14.
[202] It is no an accident that Dennett's work in *Consciousness Explained* has been variously relabeled with revised titles like 'Consciousness Denied' (see Searle, 1997) and 'Consciousness Explained Away'.

rather than *disappear*. If these processes disappear, we will clearly not have done justice to the phenomena we set out to characterise and explain.

Consciousness, Conscious Intention and Conscious Control

A crucial line of thinking that has been developing since Chapter 5 has been that, if we are to do justice to human agency in ways that best promise to make sense of and secure our claims to being free agents (and thus avoiding the threat that we might merely be AAs), then we need to foreground, clarify and defend the significance and role of consciousness, conscious awareness and conscious control. Chapter 6 focussed our attention on a particular set of challenges to the presumed role and importance of consciousness, most clearly represented by sceptical readings of Libet's data, and as expanded into a general account of the illusory nature of conscious will by Wegner.

At the conclusion of Chapter 7 it was argued that, notwithstanding the important correctives and clarifications offered in that chapter, a sceptical critic might remain unconvinced that 100-150 milliseconds could provide enough of a window for conscious processes to have any influence over behaviour, even where such processes are understood as extended processes of intentional monitoring and control occurring on an intermediate-term time scale.

Following the discussions and arguments of the present chapter, we can now see our way clear to provide an appropriately nuanced (if still incomplete) response to this objection. As I have been repeatedly emphasising, the central lesson of this chapter has been that we should *not* be looking for conscious control to somehow fit a single, simplistic model of strongly centralised and instructive control, when we have good (empirical and other) reasons to think that there will be multiple systems and sub-systems of control at work in a complex biological organism such as a human agent, including a range of dynamic and adaptive control systems that are likely to provide decentralised control over soft-assembled solutions to behavioural challenges. The role of consciousness in the conscious control of behaviour is not going to be to unnecessarily reproduce or duplicate this control by the superimposition of a centralised executive of some kind that somehow re-does the work of these other systems. Instead, we should expect consciousness to play its distinctive role *within* a larger complex system of ongoing monitoring, feedback, modulation and control of behaviour within a framework shaped, guided and constrained (as highlighted in Chapter 7) by our states of conscious intention.

With regard, then, to the question of the amount of time 'left' to consciousness for it to have its influence, as suggested by Libet's data, a number of responses are now in order. The main response motivated by the preceding discussion is that we should not be looking for signs of, and gaps for, a process of conscious control that *replicates* or fits the model of some other form of agent- or body-related control. Suppose, for example, that the RP turns out to be part of a process of decentralised control over the motor system, along the lines suggested by Thelen and Smith's (1994) account of the development of intentional reaching behaviour. We should not be looking for a system of conscious control to come along and somehow redo the work of the motor system. It would be nothing short of systemically, developmentally and evolutionarily extravagant to expect this kind of redundant superimposition. At the same time, we should not be expecting any 'activity' on the part of a system of conscious control to somehow fit the model or timeframe of these motor processes. That is, we should not be expecting that the neural processes associated with conscious control are somehow going to fit the very same rough model imposed on Libet's data: i.e. RP precedes conscious intention which precedes motor neuron activation, all over a time span of 500-700 milliseconds. Of course, if we are going to expect a 'conscious intervention' (a veto, a trigger?) to look like this, we are going to be puzzled or troubled by the apparent window of 100-150 milliseconds 'left' for consciousness to make a difference. But why should we expect conscious control and influence over intentional behaviour to 'look' like this or fit this model[203]?

It is almost as though, having mistakenly[204] expected to find a 'conscious intention' at the beginning of a causal chain leading to a given wrist flex, we must then at least expect consciousness to come along later and give an extra little nudge of intervention, and 100-150 milliseconds strikes us as inadequate to this task, especially in a context where it has been emphasised and reemphasised that consciousness typically operates on more extended timescales. But given the data and the context of the experiments, none of these voluntary wrist flexions needed any extra nudge from consciousness 100-150 milliseconds before they happened. As voluntary, intentional acts performed as part of a larger, ongoing intentional project of participating in and cooperating with the experiment, the flexions smoothly

[203] Notice that Dennett seems to make exactly this kind of (more-of-the-same) mistake in the passage quoted earlier: "This is an illusion since many of the reactions you have to that mental activity are initiated at the earlier time…" (Dennett, 2003, p.242), as if it had by now been established that all conscious activity trails after some neural activity that sets it in motion.
[204] See Chapter 7.

followed awareness that they were about to be performed. There is nothing here to make us doubt that systems of conscious control were functioning absolutely normally in all of these participants.

What about when systems of conscious control do need to 'make an intervention'? There are two immediate and short answers to this question, at least in relation to any purported relevance that Libet's data might have on the matter. First, and most bluntly, Libet's data shed *no* obvious light on what we should expect in instances when conscious intervention is required. His experiment had no such requirement, and any 'spontaneous' interventions on the part of his participants, including any so-called veto trials (where the participant claimed to have consciously vetoed the decision to flex on a particular occasion at the point when they became aware of the intention to flex), strictly speaking represented departures from the prescribed procedure for the experiment, and thus strictly speaking are bad data. Veto trials were not completed trials – there was no flexing because it was supposedly vetoed – so there could not be timing information available for us to puzzle over. There is thus no evidence to suggest that a veto does, can or must have its effects over a time-span of 100-150 milliseconds.

Second, it is exceedingly difficult to see just *how* an analogue of Libet's study could be successfully constructed to test and time conscious interventions in the course of intentional action, given the nature of precisely the interventions that need to be studied. What gave Libet's original studies some degree of methodological credibility, when it came to the issue of timing conscious intentions, was that the study required the participant to be essentially passive – the 'wish' or 'intention' to flex should just be allowed to arise spontaneously, such that the participant only had to divide their attention between monitoring their consciousness of their 'intentions' and the position of the oscilloscope 'clock'. But real-world examples of contextually rich and meaningful intentional action in which we identify some particular instance of (assumed) conscious intervention and control can only have their validity compromised by trying to divide attention between the (assumed) intervention and some other exercise in timing consciousness.

Once again, I repeat the message of this chapter as a whole. Control in a complex biological system like a human agent is not going to just look like one thing, or just neatly fit one model. But then why should we expect to find that conscious control will look just like one

thing, or fit one model, let alone replace, supplant or (unnecessarily) replicate these other systems of control? The answer is that we should not. Instead, we should expect consciousness to play its role within a complex ongoing process of behavioural initiation, monitoring, feedback and modulation in the light of conscious intention. Whether or not there might be some more distinctive form of conscious control is something that still needs to be addressed in the chapters to follow. But Libet's studies certainly do not raise issues about the need or form of such control.

Here is one example of the kind of real-world, contextually rich and ecologically valid behaviour that would be an appropriate test case to see conscious control mechanisms at work. Consider the remarkable feats of intentional behavioural control achieved by a concert pianist in performance. With ten fingers plus feet at work, it is obvious that a pianist's conscious control over their performance does not lie at the level of detailed motor instructions to individual body parts. Through extensive practice, these sequences of movements have been automatised over time, leaving the pianist to focus most of their conscious attention on the performance as a whole, especially on its more expressive, emotional dimensions. At the same time, the pianist is free to shift their attention from the flow of the performance as a whole down to the timing and force of a particular note. Consciousness can 'intervene' to influence this individual note, or intervene to adjust the pace of passage in which little or no attention is allocated to the details of the automatised motor sequence.

This is what consciousness looks like at work in the real world, and it is a picture that (I have argued) is not threatened by the data that emerged from Libet's laboratory. It is also a picture that points to the more ordinary, agency-enhancing aspects of our capacity for automatising elements and sequences in various skilled performances, in contrast to a Bargh-like view of automaticity as somehow undermining our agency and control over our lives[205]. Finally, it is a sketch of the conscious agent in action in which the alleged illusionary status of conscious will looks like an implausible box-and-arrow abstraction that no-one could sensibly apply to the dynamic flow of the unfolding performance[206].

[205] We will revisit these aspects of learning and skill acquisition in Chapter 10.

[206] Where, we might ask, would Wegner like us to begin disaggregating the myriad elements of this performance in order to impose the schema outlined in Figure 6.1.? There is no part of this dynamic display of embodied control that can be isolated from the rest as though it were somehow a discrete action about which we have the illusion of its having been initiated by an equally discrete conscious intention.

It is certainly not the case that we understand all or even most of the details of how we achieve these feats; nor have we have solved all the pressing problems it is possible to raise about consciousness. Nevertheless, a suitably nuanced view of the matter, both with respect to important conceptual and philosophical issues (Chapter 7) and to important empirical clarifications as to the nature and diversity of systems of control in biological systems (this chapter), provides us with sufficient justification to resist the more sceptical interpretations of those data.

Nevertheless, an as-yet unsatisfied critic might want to push one more objection. Suppose it is conceded, for the sake of the argument, that the central conclusion of this chapter is correct – that we should not expect control in a human organism to just fit one model. Suppose it is further conceded that we can stop the slide towards a Dennett-like dismantling of the agent – we can successfully 'smear' human agency in time and space without advocating excessive scepticism about consciousness and the conscious self. In the context of a project intent on defending *free* agency, can we not still ask the following question: are the models of control described in this chapter, and the resultant complex interplay of control systems within a human agent that is envisaged, any more *friendly to freedom* than the strongly centralised, implicitly instructional model of control my arguments seem intent on destabilising, if not displacing? Perhaps the latter model is more vulnerable to the apparent threats posed by the data and arguments of Chapter 6; but perhaps it can and should be defended rather than abandoned, especially if it promises to be more freedom-friendly than the decentralised and distributed alternatives?

I would first make two observations about this objection. First, the arguments and discussions in the current chapter have pointed to important questions of *empirical plausibility* when it comes to the assumptions and claims we make about control in biological systems. If Chapter 7 suggested that the model/s of agency assumed in the sceptical arguments of Chapter 6 were conceptually and philosophically problematic, our current discussions add a weight of empirically-based argument against assuming or proposing such models. To reject (or ignore) these arguments because it is as yet unclear to what extent the alternatives are freedom-friendly would be to miss the point. The alternative models demand attention because their empirical credentials must be weighed against the consequent empirical implausibility of the more traditional model.

Second, the objection is somewhat premature. On one hand, we have yet to consider in any great detail a model of distinctively human control in agency, and it is thus premature to ask whether the role/s given to more decentralised and distributed systems of control within that model are freedom friendly or not. On the other hand, the proposals that will be developed here can only hope to be programmatic – that is, they will be proposals that point to promising avenues along which we might look to locate and explicate the grounds of human free agency. To ask of such programmatic proposals whether or not each and every element of the sketch of agency on offer makes a freedom-friendly contribution is, again, rather premature.

Be that as it may, I have two brief responses to offer our critic. The first is that, if your preferred model of agency (and, thereby, of free agency) is one involving subjugation to the demands and strictures of the rational will[207], then I suppose the model (or, more modestly, the sketch) of agency emerging here will look less freedom friendly than a more centralised, instructive alternative. But, as we are about to see in Chapter 9, there are strong empirical and philosophical considerations that argue *against* assuming too much about the powers and reach of the rational will. To state the case in the terms introduced in Chapter 5, I will argue that we need to avoid claiming that we are, or that we should aspire to be, *Hyper-reflective Hyper-rational Agents* (*HHA*s) if we want our defence of human free agency to pass both empirical and philosophical muster.

My second response is that, as has been evident in the preceding sections of this chapter and elsewhere, a significant part of the current project (and programmatic proposal) is to foreground *consciousness* and *conscious processes* in human free agency. In this context, questions about occurrent processes, capacity and capacity limitations take on a significance that tends to be neglected in more traditional reason-based accounts of deliberation, choice and action[208]. Relative to this context, the models of control discussed in the current chapter can (and ought to) be exploited in order to give conscious processing its proper place in our

[207] That is, for example, something along the lines of Mele's (1995, 2006) *ideally self-controlled agent* whom we have encountered in earlier chapters.

[208] See my discussion of the tendency to avoid commitments regarding consciousness and occurent mentation in Chapter 5. It is not that questions of ability and capacity are irrelevant on reason-based accounts – Mele's ideally self-controlled agent requires many important abilities. It is rather than that these tend, like reasons and propositional attitudes themselves, to be less obviously related to occurent mental processes and the capacity limitations we might associate with these.

account of agency by, in part, reserving conscious capacity so it can be utilised where it is most needed[209]. In this sense, the models of control discussed in this chapter are importantly *freedom neutral*; but they promise to make a contribution to an empirically-informed, empirically- and philosophically-plausible, and consciousness-friendly account of human free agency.

There is, of course, much more to say about consciousness and its importance to free agency. Specifically, more needs to be said about the way/s in which consciousness contributes to a form of self-governance that we might consider distinctly human. In the following chapter, I will take this project forward by, paradoxically, sounding some notes of caution about just what we think the reflective, monitoring and reasoning powers of conscious thought could amount to.

[209] For example, in the monitoring and modulation of a performance, as with our pianist, rather than in the (extravagant) initiation of a myriad micro-actions, each one requiring the formulation of a conscious intention.

Chapter 9

Realistic Self-governance I: Consciousness, Emotion, and the Limits of Reflective Deliberation

Thus far in this thesis, consciousness has been taking on an increasingly prominent role following my proposal of a framework for investigating agency and freedom that is no longer defined by the overarching concerns of the traditional free will debate between compatibilists and incompatibilists. In Chapter 5, I argued (in partial agreement with O'Connor) that giving a distinctive role to consciousness was a critical component of any attempt to defend claims of free agency against the threat that humans might, in the end, turn out to be agent automatons (AAs) of some kind. And in Chapter 6, we saw how the stakes have been raised in this regard by the accumulation of a range of empirical evidence that has been used to mount various sceptical arguments against the significance, role and reach of consciousness in the initiation and control of human behaviour.

In Chapters 7 and 8, much of this ground has been progressively clawed back from the grasp of the sceptics. Chapter 7 highlighted and defended a number of philosophical correctives and qualifications that together undermined crucial assumptions about (*inter alia*) agency, intentionality and conscious intention that are at work in certain of the sceptical arguments in Chapter 6; and in Chapter 8, an empirically motivated corrective was proposed and defended regarding our expectations about the kind/s of control systems we might expect to find at work in complex biological systems like human organisms; and implications of this corrective for how we might reasonably expect conscious control mechanisms to feature in human agency were similarly proposed and defended.

While these discussions and arguments have not been negative *per se*, they do represent more of an attempt to defend a *space* for human consciousness to make a distinctive contribution to human agency, and thereby to free agency, without necessarily filling in all that much detail as to what precisely that/these contribution/s might be. The challenge to be confronted in this and the following chapter is thus one of trying to fill in some more of these details. This is a substantial project in its own right, and thus one on which I can only hope to make a constructive start that points to promising avenues for future research.

Given the size of the positive project involved, it is sensible to start with some candidate ideas that have, as it were, done the rounds in philosophy's on-and-off engagement with the promise and perplexities of consciousness. The obvious starting place, given our overarching concern with questions about agency, is with variations on the idea that the distinctive contribution of human consciousness comes by its association with and role in processes of self-reflection, interpretation, reflective endorsement, deliberation, as well as in recognising, imposing and complying with normative demands and constraints on behaviour. Although something of a loose collective of ideas, these together represent a strong and influential tradition in Western thought of prizing our status as rational agents endowed with powers of reflection and insight that allow us to explicitly (hence consciously) think our way towards the 'best' course of action, given our insight into not only the ways of the world but also into our own characters and interests.

There is a huge potential pool of examples in which we can find a link being drawn between our capacity for reflective and deliberative consciousness and our special status as rational agents. My sample below is simply intended to be representative of a range of views that illustrate some of the ways and contexts in which this kind of link has been made.

The Kantian tradition is strongly associated with the interweaving of reflective and deliberative consciousness with our status as rational agents. In *The Sources of Normativity*, Christine Korsgaard (1996) offers us her own neo-Kantian take on the intimate connection between the structure of and capacity for reflective and deliberative consciousness, and the origins and grounds of our capacity for normativity and autonomy. Her views nicely illustrate the way in which the collection of ideas listed above can be seen at work in accounts of rational and autonomous agency. First, we can see Korsgaard making explicit the link she sees between reflective self-awareness and our being able to take up what we can call a deliberative position with regard to our actions, and from which we can legislate to ourselves:

> The reflective structure of the mind is a source of `self-consciousness' because it forces us to have a conception of ourselves...When you deliberate, it is as if there were something over and above all your desires, something which is you, and which chooses which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of yourself. (Korsgaard, 1996, p.100)

Ultimately, for Korsgaard, this leads us to the idea of the agent as a self-legislator – as being a law unto themselves. I will not follow the detail of the argument here, but in crude outline it

works something like this: (i) the reflective structure of human consciousness forces upon us a conception of ourselves as agents – a self-conscious identity; (ii) the structure of reflective consciousness in deliberation is also such that we must first endorse our desires before acting on them – we must turn them into reasons for acting; (iii) in 'forcing' on us a conception of ourselves as something over an above all our particular (contingent) desires, the self-conception that remains available for us to identify ourselves with in the deliberative position is our identity as rational agents; (iv) but if we are to find reasons to act consistent with our identity as a rational agent *qua* rational, without any particular contingent desires or other aspects of character to ground those reasons, then these must be reasons that are valid for any rational agent *qua* rational; thus (v) the reflective structure of deliberative consciousness ultimately forces on us an identity and a perspective of, as it were, bare rational agency from which our reasons for acting are as laws for all rational agents (the idea of the Categorical Imperative). Korsgaard concludes her argument thus:

> The reflective structure of consciousness requires that you identify yourself with some law or principle which will govern your choices. It requires you to be a law to yourself. And that is the source of normativity...[O]ur autonomy is the source of obligation. (Korsgaard, 1996, p.104)

Of course, the details of Korsgaard's argument, and of her larger project, matter a great deal to the plausibility of this argument; and there is a whole lot of conceptual baggage at work here from, especially, Kant's views on ethics and practical reason. But it does provide a very clear and strong example of the way in which ideas about the structure of human reflective and deliberative consciousness have been taken to mark out a distinctive domain of cognitive, deliberative and normative activity that identifies us most deeply as reflective, deliberating, rational agents. Dramatic things are claimed to follow from our capacity for reflective and deliberative consciousness, and those things are most closely associated with reasons, rationality, and a kind of reflective distance from our (contingent) profile of interests.

While Kant may have considered compatibilism a "wretched subterfuge"[210], his ideas about autonomy and the need to endorse our desires before acting on them have obviously been influential within compatibilism. In Chapter 3, I surveyed various compatibilist contributions to the understanding of free agency, with a special emphasis on the importance attached to ideas of reflective endorsement. I will not rehearse these ideas here. Suffice it to say that, however scant some accounts may be regarding the occurrent processes associated with

[210] Kant (1788/2010) – see my discussion of Kant's complaint in Chapter 1.

reflective endorsement, it is fair to posit conscious reflection, self-interpretation and self-attribution, and conscious higher-order volition and endorsement/ alignment as crucial features of these accounts when fleshed out in greater empirical detail.

Even Dennett (2003), who is (as we have been reminded) hardly a friend of consciousness, ultimately locates the distinctive mark of human agency in the area of giving and demanding reasons for action. With respect to our more distinctive capacities for agency, Dennett (2003) locates the critical evolutionary break between humans (or our ancestors – he does not specify) and other creatures at the point where our (apparently unique) communicative capacities emerge. Citing the work of David McFarland (1989, in Dennett, 2003), Dennett links the emergence of new capacities for self-monitoring to communication as follows:

> It is only once a creature begins to develop the activity of communication, and in particular the communication of its actions and plans, that it has to have some capacity for monitoring not just the results of its actions, but of its prior evaluations and formation of intentions as well. (Dennett, 2003, p.248)

In this evolving social context of communicating actions and plans between individuals, Dennett imagines the emergence of a user-interface (analogous to the software user-interface on a computer), centred around the user-illusion of 'the self', that provides both others and ourselves with a useful but limited perspective on the attitudes, evaluations, plans and purposes lying behind behaviour.

On Dennett's (2003) account, then, the push towards increased monitoring and control of individual behaviour is external and social, driven by the emerging activity of communication – where communication both facilitates coordination of plans and behaviour, whilst also creating new possibilities for competition and manipulation. Because of this link to communication, the 'self' associated with this self-monitoring activity is fundamentally grounded in and mediated by systems of communication – it is a linguistically mediated practice – driven by the demands of a novel social-evolutionary context: "…we wouldn't exist [as selves]…if it weren't for the evolution of social interactions requiring each human animal to create within itself a subsystem designed for interacting with others" (Dennett, 2003, p249). And as part of this emerging communicative practice, we began to give and demand reasons of each other, in the process "bootstrapping ourselves to freedom."[211]

[211] The latter phrase comes from the title of Dennett's chapter 9 in *Freedom Evolves*.

This is evidently an outside-in story of how our distinctive capacities for self-reflection, self-monitoring, and reason-based deliberation took shape, and one that does not associate these latter capacities with consciousness *per se* (or better, since the account is coming from Dennett, it describes processes that most of use would associate with consciousness, and that Dennett may well be happy to call conscious, except that he would not want that to have any special implications about the kind of awareness involved, or any 'phenomenal experience' we might think of as marking out these processes as distinctive). But in the end it is a story that amounts to a view of reflective and deliberative consciousness as lying at the heart of distinctively human capacities that ground and shape our agency and our claims to freedom.

As a final example to add to this small but useful sample, we can include Gallagher's view of conscious action and free will that we have encountered at various points so far (see Chapters 7 and 8). For Gallagher, the appropriate level at which to ask questions about consciousness and free will is the extended timeframe in which a reflective and interpretive consciousness becomes involved in the ongoing cycles of feedback loops that characterise our behavioural forays in the world. So, in the example of a chance encounter with a snake cited earlier (Chapter 7), Gallagher sees consciousness as being crucially involved on the time scale of seconds (following the earlier quick and automatic fear and startle response) where the situation gets explicitly and reflectively interpreted, the snake is consciously perceived and identified (interpreted) as harmless, and is approached with a view to pick it up based on reflective awareness of an interest in doing so. The hallmarks of distinctively human agency are thus, for Gallagher, the interpretation of experience and action, and the taking up, pursuit and sustaining of intentions and intentional projects, where these processes cannot be conceived of as occurring in the absence of a suitable reflective, interpretive and deliberative consciousness. In *How the Body Shapes the Mind*, he concludes his discussion of Libet and free will with this summary of his take on the role of consciousness:

> In this complex interaction [of agent and environment] conscious decision-making – the taking up of intentions – the interpretations of what we experience – can shift the system [i.e. the agent] and alter the biases [affecting the system], can create new biases that in the long run add up to 'character' – which in turn may determine future responses. (Gallagher, 2005, p.243)

There is much in the above-cited views that deserves fuller discussion, and we will return to some of the ideas in due course. But my immediate intention is not to extract and defend what might be right in each of these views, while jettisoning the more problematic bits. Instead, I want to begin a move towards a more positive account by asking a different kind of question

raised by these views, and foreshadowed by my discussion of the idea or image of the hyper-reflective hyper-rational agent (HHA) in Chapter 5.

Let us call the perspective on what is distinctive about human agency that runs through these (and other) accounts a reflective deliberation (RD) view of what is most distinctive about human consciousness and its contribution to agency, including free agency. The question I want to begin by asking is this: does an RD view describe a type of consciousness (and thus a form of conscious agency) that we can realistically claim on behalf of the 'average' or 'normal' human agent? That is, does the RD view of the human agent give us an empirically plausible view of human agency, such that there are no obvious obstacles to asserting that the vast majority mere mortals qualify as possessing and manifesting this form of agency?

Closely related to this question of empirical plausibility is a more obviously normative question: is the form of consciousness and agency described on the RD view normatively desirable? That is, whatever the long tradition of esteem that has been attached to our capacities for reflection and rational deliberation, does the RD view outline a form of consciousness and agency that we ought to aspire to? Is RD consciousness something we should aspire to as an ideal? And is it something which, *qua* normative ideal and ground of our most distinctive capacities as agents, we ought to in some sense *maximise*, such that we become *better* or *more effective* agents, or agents possessed of greater *freedom*?

I cannot hope to offer definitive answers to either of these questions in what remains of this thesis. I do want to make two important contributions to answering these questions, however. First, I want to provide a number of compelling reasons for urging caution in assuming or adopting too strong a view of the reflective and deliberative capacities of human agents. These reasons include: (i) evidence for limitations in our capacity for self-insight; (ii) evidence that too much reflection, too much thinking about what we are doing, too much reflective detachment, are not necessarily a good things for agents like us; and (iii) evidence that a particular form of reflective distance, or reflective detachment, specifically with regard to the relationship between reason and emotion, both mischaracterises the role played by emotion in our practical reasoning, but also implicitly prescribes a normatively undesirable deliberative position for human agents. The evidence and arguments relating to (i) to (iii) will form the substance of the rest of this chapter.

More positively, and partly in the light of these notes of caution, I also want to propose and argue for the importance of two less-well explored avenues in which we might find or locate distinctive degrees of freedom in human agency besides the more obviously reason-based deliberative aspects. These avenues are (a) the imagination and (b) externalised aspects of mind. The details of these proposals will be the subject of Chapter 10.

Realistic Constraints on Conscious RD Agency: Self-insight

A good place to start in noting important constraints and limits on our conscious agency is with the material presented in Chapter 6 that has so far escaped contention. If there is one thing that is established quite clearly by many of the studies cited by Bargh and colleagues, as well as by Wegner, in their various attempts to bring consciousness 'down to size', it is that we are quite capable of misinterpreting our own behaviour and its relation to our intentions and other mental states. Further, we are quite capable of doing things of which we almost entirely unaware. Whatever our remarkable capacities for reflection, self-insight and action with conscious intent, we are most definitely fallible in this regard.

It will be helpful to recall some of the examples of the odd things that people do – or that they can be manipulated or tricked into doing – at least as demonstrated in various laboratory-type scenarios (whose internal validity we will not question). Bargh and colleagues catalogue the following cases from the automaticity literature:

- Without any awareness of the influence, people exposed to words connoting rudeness will (comparatively) act more rudely themselves.
- Without any awareness that a stereotype of the elderly has been primed or activated, people (unintentionally) adjust the way they walk and how much they recall from an earlier part of the experiment, in ways that are consistent with what an old person might do.
- People appear susceptible to having various cognitive goals activated without them being aware of this, such that they will behave similarly to others given explicit instructions to pursue a certain cognitive goal without any awareness that this was what they were doing.
- People can have behavioural goals, such as an achievement goal, activated without their awareness, such that they behave as if they were pursuing this goal, and react to success or failure as if achievement mattered to them, while they profess no awareness that this was what they were doing.

- A range of evidence suggests that people's affective and evaluative responses are primarily automatic and engaged independently of conscious awareness, appraisal and deliberation. So, for example, an individual provided with a subliminal prime for an object or situation they are likely to evaluate positively (or negatively) displays quicker response times for positive (or negative) adjectives on a subsequent task, suggesting that they automatically evaluated the primed object or situation. Both the processing of the prime, and the apparent evaluation carried out, bypass conscious awareness.

Wegner seems happy to accept the automaticity examples as grist to his theoretical mill, and adds a number of other apparent quirks, failures and shortcomings of our capacity for conscious self-insight and agency:

- Evidence of demonstrably intentional action that is *not experienced or claimed* as having been intentionally caused. In addition to examples from the automaticity literature, Wegner adds phenomena like 19th century table-turning as relevant cases here.
- Evidence of cases where conscious intending or willing of an action is experienced or claimed by an agent under circumstances where the candidate 'action' was demonstrably *not intended* by the agent. Here, Wegner reports that the right combination of circumstances (as regards causal self-attribution) can lead participants in laboratory studies to report intentional behaviour (stopping a cursor at a particular point on the screen) when the behaviour is demonstrably under the intentional control of another (the researcher's confederate).

One could raise philosophical quibbles about much of this evidence, especially in terms of the precise sense in which the automaticity cases, and Wegner's intentional-without-awareness-of-intention cases, represent intentional behaviour. But such quibbles and clarifications are peripheral to the significance of this evidence to the current topic, which involves the degree and accuracy of our self-insight into our intentions, goals, reactions, behaviours, and actions. It is not conceding too much to the likes of Bargh and Wegner to admit that their evidence demonstrates, at the very least, that we are both limited and fallible when it comes to self-insight, and the attributions and explanations we offer for our own state of mind and our behaviour. We do not have infallible access to, or infallible knowledge of, these features of ourselves and our activity.

This modesty about our capacity for self-insight is strongly reinforced in a long line of literature on (non-pathological) confabulation. As Peter Carruthers puts it in his recent work on mindreading and introspection:

> There is extensive and long-standing evidence from cognitive and social psychology that people will (falsely) confabulate attributions of judgments and decisions to themselves in a wide range of circumstances, while being under the impression that they are introspecting. (Carruthers, 2009, p.130)

Included in the research that Carruthers cites here is classic research by Nisbett and Wilson (1977), alongside research conducted by Wegner, including the computer cursor study described in Chapter 6 (Wegner & Wheatley, 1999). For Carruthers (2009), this evidence of widespread confabulation is consistent with, indeed provides support for, his claim that first person metacognition (including reflection and deliberation over attitudes and judgements) depends on what are typically third-person processes of mind reading and attitude attribution directed at oneself, not on introspective access to our propositional attitudes.

Without engaging the details of Carruther's (2009) argument, or trying to resist its full force, it seems safe to assume that the evidence for confabulation that he highlights at least successfully undermines and strong views on transparency, privileged access, and immunity to error when it comes to the attitudinal states most closely associated with traditional accounts of rational reflection, deliberation and choice.

Realistic Constraints on Conscious RD Agency: The Deleterious Effects of Thinking Too Much

At the conclusion of Chapter 8, I made use of the example of a concert pianist shifting their attention and interventions between the overall shape and feel of the performance, and the emphasis and timing of a single note. This example was meant to illustrate the power and flexibility of conscious control in action in richly contextualised real-world activity. The same example can, however, just as well be used to illustrate the possible downside of too much conscious reflection, too much self-consciousness, and generally the downside of thinking too much.

In commenting on the importance of automaticities to phenomena such as highly skilled performance, the psychologist Merlin Donald highlights just these sorts of potential

downsides to trying to have or keep too much ‘in’ consciousness:

> We can carry on several parallel lines of cognition at the same time, provided they are kept out of consciousness. Musicians know this. When professional pianists play, they cannot afford to become overly conscious of their fingering or the specific notes of the passage they are playing, particularly the more rapid ones. That kind of self-consciousness is paralyzing. They have to automatize those difficult passages, or they will make major mistakes. The same rule applies to speaking. (Donald, 2001, p.26)

Indeed, the same rule arguably applies to most of our waking activities. On one hand, we cannot afford the effort and the inefficiency of trying to direct our consciousness at the details of our bodily implementation of actions. When I interrupt my work to fetch a particular text from my study, my consciousness is better directed at keeping in mind my intention (lest I find myself in my study with no idea why I am there – an experience we are all familiar with) than on the individual steps I must take in order to get there. Or, moving around in an unfamiliar environment, I am better off directing my attention to obstacles and potential obstacles than to the step-by-step details of my perambulations. On the other hand, we often cannot afford the consequences that follow from heightened self-consciousness of ourselves in action, as opposed to on the world and the context in which we act. Better that I be focussed on the content and flow of my lecture, and on my audience, than that I be self-conscious of my choice of words, my delivery, my appearance, and what my audience might be thinking about all of these.

These kinds of limitations on the value and efficacy of conscious reflection and self-monitoring are nicely illustrated by a phenomenon that psychologists have labelled ‘verbal overshadowing’. As the label suggests, verbal overshadowing is a phenomenon specifically involving language. It denotes cases where someone’s verbalising what they are doing has a detrimental effect on their performance of the task related to the verbalised material (Chin & Schooler, 2008). According to Chin and Schooler (2008), verbal overshadowing effects have been noted in face recognition, decision making, problem solving, analogical reasoning, and visual imagery. While the causes or sources of verbal shadowing effects remain controversial (Chin & Schooler, 2008), the general idea of this kind of effect of thinking too much, of trying to reflect and make things explicit to oneself (and others) with the effect that performance is impaired, illustrates nicely the kinds of constraints that accounts of RD consciousness and conscious agency need to accommodate.

Documented cases of the verbal overshadowing effect range from the troubling to the amusing. In the former category, the original study in which verbal overshadowing was first

named as a phenomenon (Schooler & Engstler-Schooler, 1990) focussed on issues of eye-witness face recognition and memory that might be relevant to later identification of suspects. Having watched a video of a robbery, participants in the verbalising condition were asked to describe the robber, while those in the control condition engaged in an unrelated reading task for a similar amount of time. The participants who had verbalised a description of the culprit were found to be significantly worse at picking the robber from an identification line-up than were the controls.

On the more amusing side of the phenomenon, Melcher and Schooler (1996) reported the results of a study on the effects of verbal overshadowing on wine tasting and recognition. Melcher and Schooler (1996) grouped their participants into those who were non-wine-drinkers, those who drank wine but had no trained expertise in wine tasting, and trained expert wine tasters. Each participant tasted a red wine, and then either engaged in a task requiring verbal description of the wine or in an unrelated verbal task. Participants then had to try identify the wine they had tasted earlier from a group of three wines. While verbalisation seemed to have little to no effect on recognition in the experts and the non-wine-drinkers, the untrained wine drinkers showed a clear tendency towards a verbal overshadowing effect, with their ability to recognise the wine they had tasted being significantly impaired. This was especially noteworthy in the light of the fact that untrained wine drinkers who had not verbalised a description of the wine were almost as good at recognising the wine they had tasted as were the experts from either condition.

There are reasons to be cautious before trying to read too much about consciousness and conscious reflection into these results. Many of the demonstrations of verbal overshadowing, including most of the original studies, fundamentally involve memory, and thus seem to be, in the first instance, about the role of linguistic encoding in enhancing or impairing memory for non-linguistic stimuli. Exactly how this might have relevance to general issues of occurrent conscious processes, including self-reflection, insight and deliberation, is not immediately obvious.

However, Melcher and Schooler (2004) do suggest a broader conception of verbal overshadowing that extends beyond applications to just memory and recognition. They describe verbal overshadowing as involving the “disruptive effects of articulating nonverbal cognition in an array of domains, including perception, memory, and problem solving”

(Melcher & Schooler, 2004, p.618); and they cite a number of studies from domains of cognitive activity other than memory *per se*, including affective decision making, insight problem solving, visual reasoning, and analogical transfer.

One study in particular, by Wilson and Schooler (1991), takes us to the heart of the issue at stake in this section. Titled "Thinking Too Much", the paper presents the results of two experiments which, to varying degrees, suggest that analysing, or attempting to introspect, the reasons for one's preferences in a decision-making context may impair the quality of the relevant preference judgements and related decisions. For example, in their first experiment, Wilson and Schooler (1991) had two groups of college students rate their liking of five different brands of strawberry jams, and compared these ratings to those of a panel of expert tasters[212]. The control group was not did not receive any special instructions besides those relating to the rating how much they liked each jam, whereas the experimental group were instructed to "analyze why you feel the way you do about each jam, in order to prepare yourself for your evaluations" (Wilson & Schooler, 1991, p183). They were further told that they would have to list their reasons for their liking/disliking ratings after they had tasted the jams.

There was a significant degree of agreement between the ratings of the control participants and the expert judges, but not between the reason analysers and the judges ratings. The reason analysers cited reasons for their liking ratings that did not agree with expert opinion; while the degree of liking for each jam implied by the reasons offered correlated strongly with the final preference ratings given by these participants. Wilson and Schooler (1991) concluded that, for these reason analysing participants, being asked to think about why they did or did not like each of the jams had brought to mind reasons that did not match very well with the opinions of experts such that, having subsequently based their preference ratings on these reasons, these did not match the ratings of the experts terribly well either. Since the controls showed no such systematic divergence from the experts, Wilson and Schooler's (1991) interpretation is essentially as follows: (a) both controls and reason analysing participants would, on tasting each jam, have had roughly the same order of preferences as the experts, based on a number of sensory characteristics (individually rated by the experts); but (b) when the reason analysers came up with reasons for liking or not liking each jam,

[212] The five jams had been respectively rated 1st, 11th, 24th, 32nd and 44th by an expert panel in a previously published report (Wilson & Schooler, 1991).

citing various characteristics, their reasons did not correspond to the experts' opinions; and (c) having thus (consciously) focussed in on the 'wrong' characteristics, articulating these reasons effectively changed these participants' minds about how they liked the jams, thus producing the discrepancy with the control and expert ratings.

Wilson and Schooler are careful to point out that they do not think introspection and reflection on reasons for decisions will somehow always yield "nonoptimal choices" (Wilson & Schooler, 1991, p.191). The deleterious effects of 'thinking too much' might well vary according to the quality/accuracy of our initial feelings or preferences, according to how knowledgeable we are about the attitude object, and according to the time available for either a shallower or a deeper analysis. Nevertheless, Wilson and Schooler (1991, p.181) take it that they have provided solid evidence for the claim that "analyzing reasons can focus people's attention on nonoptimal criteria, causing them to base their subsequent choices on these criteria."

There is thus a kind of double significance to evidence for verbal overshadowing effects. First, the phenomenon establishes that there are a range of contexts in which thinking too much, reflecting too explicitly on what one is doing, and specifically doing this by trying to articulate aspects of one's mental state and/or performance, can be deleterious to that performance. Second, as if to illustrate and reinforce the observation that humans are at risk of becoming what they believe themselves to be, even if what they believe is wrong[213], the evidence also suggests that once we have managed to interfere with our performance (as in the case of identifying and articulating "non-optimal criteria" for judging the jams), we then compound the error by adjusting our judgements (in this case, of the jams) to fit our explicit criteria rather than our (original) response.

This is neither evidence nor argument against the possibility of self-knowledge, nor against the potential value of explicit reflection on and articulation of thoughts, judgements, and activities[214]. Nevertheless, as was the case in the previous section of this chapter, the evidence does at least suggest some serious constraints on both the empirical possibilities for,

[213] "When reinterpretations of ourselves are taken seriously, they not only have the power to change our view of others for the worse, but even more power to change our own self-definition, so that we start to live up to them." (Blackburn, 1998, p153)

[214] See Wilson (2009) for his own view on the importance, value and limits of self knowledge. See also Wilson and Dunn (2004).

and the adaptive and normative desirability of, an overly strong emphasis on reflective and deliberative aspects of conscious agency.

Realistic Constraints on Conscious RD Agency: Reason, Emotion & Detachment

The third source of constraints I want to highlight represents a mix of empirical and philosophical concerns about the relationship between reason and emotion, or more specifically, about the proper role of emotion in deliberation and choice. The primary motivation for raising this issue comes from empirical observations in the fields of neuropsychology and neurology of patients with damage to the ventromedial areas of the prefrontal cortex. What seems remarkable about these patients, from the point of view of trying to understand 'normal' agency, deliberation and decision-making, is the peculiar combination of deficits and (relatively) preserved functions that they display, and most especially the apparent dissociation that their cases suggest between certain forms of intellectual functioning, on one hand, and effective real-world agency on the other.

The key source for this evidence comes from the influential work of Antonio Damasio in his book *Descartes' Error* (1994); but Damasio's ideas are also usefully and, for the most part, enthusiastically brought to the attention of a more philosophical audience via the work of Simon Blackburn. In *Ruling Passions*, Blackburn (1998) notes the same critical point of interest mentioned above in his discussion of Damasio's rich case material:

> Damasio presents patients whose capacities for attention, whose intelligence, memory, abilities to hold several things in mind at once, social knowledge, moral reasoning, inferential or logical abilities, and language, are quite intact, yet whose capacity to live their lives in any sensible way is virtually zero. What has been lost is the application of such knowledge, in the formation of emotional affect, and thence in decision-making. So such a subject can say which of two alternatives is better, what the consequences of one or another would be, and select and register (verbally) those aspects of the situation that people count as important considerations. But all this fails to translate into action. (Blackburn, 1998, p125)

As intimated above, what should be most striking these case studies of Damasio's (1994) is that patients with this pattern of damage to the ventromedial prefrontal cortex (VMPFC) appear to have at their disposal the kinds of capacities we would most obviously associate with an intact capacity for rational reflection, self-insight and deliberation – the ability to infer consequences from alternative courses of action, to point out (at least verbally and, one might say, at a level of intellectual insight or understanding) salient features of decision-making situations that are important considerations, and even to make apparent judgements about alternatives being better or worse. From a philosophical point of view, where intellect,

systematic reflection, logical inference and cost-benefit analyses would appear to be amongst the most important building blocks of a conscious reflective deliberative agency, one would expect patients of this kind to be, at worst, only minimally impaired when it comes to decision-making. The cases, however, suggest otherwise.

The claim made by Damasio (1994) is that preservation of these capacities is of little benefit to the patient as a would-be effective agent because what has been lost is the capacity for appropriate links and interactions between the more cognitive/ intellectual systems in the individual and their systems for affect or emotion. Specifically, these patients experience a disruption or, *in extremis*, an absence of 'normal' associations between features of the decision-making context that are noted or represented or registered by the agent, and affect or emotion that would normally be attached to, or would normally mark or flag those features.

In addition to the historically famous case of Phineas Gage, the central case study presented by Damasio (1994) is that of one of his own patients named Elliot. Elliot had been diagnosed with rapidly growing meningioma located above the roof of his eye sockets, and thus pressing upwards into the medial portions of his frontal lobes. By the time the tumour was surgically removed, it had grown to the size of a small orange, and the areas of his frontal lobes damaged by the tumour were also removed. Elliot's prognosis seemed promising, and his physical recovery was swift. He also appeared to show no signs of language or obvious cognitive difficulties. But unfortunately for Elliot and his family, his life was about to fall apart.

Despite the apparent preservation of his premorbid skills and knowledge, Elliot was soon to lose his job because he was no longer capable of being left to complete a task on his own. As Damasio (1994) describes it, Elliot would vacillate between perseverating on one aspect of a task when the overarching goal required that he switch to a different aspect or activity, or switching to an unrelated activity that momentarily gripped his attention when his larger task required that he stick to what he was doing. He might never get going with a task, instead endlessly deliberating over the possible ways in which the job could be done without ever choosing one; or he could begin a task and immediately get distracted and bogged down in a peripheral activity without ever getting back to the job at hand.

His decision-making in his life was quite generally impaired, including making ill-fated business decisions against which he had been more than adequately advised. Bankruptcy was followed by divorce, his wife and children having had enough of trying to cope with the apparent irrational choices of an otherwise intelligent man. This was followed by another brief and ill-advised marriage, another divorce, and eventual dependence in the custody of a sibling where, prior to Damasio's investigations and intervention, Elliot's benefits were denied for a time on the advice of experts who, in the light of his preserved intellectual functions, saw his behaviour as either malingering or laziness.

Damasio (1994) reports almost immediately noting an apparent absence of emotion and emotional responsiveness in Elliot. Elliot could relate the sad tale of his change in circumstances without any hint of sadness, anger or frustration; and he could be exposed to emotionally provocative stimuli without showing any sign of affective shift, even while he could cognitively or intellectually acknowledge the content of these stimuli – a state Damasio evocatively describes as "*to know but not to feel*" (Damasio, 1994, p.45, italics in original). Moreover, Elliot seemed aware that his reactions to situations had changed considerably from before to after his illness and operation. Damasio suspected that this apparent change in Elliot's capacity for emotion – his apparent 'cold-bloodedness' following his surgery – lay at the root of the latter's decision-making difficulties.

It is important to emphasise, in the context of this chapter, the full extent to which Elliot's cognitive and intellectual capacities that we might regard as central to successful deliberation and decision remained intact. Damasio (1994) relays the results of a number of tasks and tests in which Elliot displayed average/ normal or superior performance[215], including (a) generating options for action in social situations, (b) awareness of, and spontaneous inclination to consider, the consequences of action, (c) means-end problem solving in terms of conceptualising effective means for achieving a range of social goals, (d) prediction of social consequences, in terms of the most likely outcome (from a range of possibilities) of various interpersonal situations, and (e) moral reasoning and judgement, involving generating solutions to moral dilemmas plus ethical justifications for the proffered solution[216]. Given his

[215] Details of the testing and performance of Elliot were first reported in Saver and Damasio (1991), where Elliot was referred to as patient EVR. It is remarkable to examine in detail the range of tests carried out, and the number of 'superior' scores obtained by Elliot.

[216] Damasio (1994) reports that Elliot's score on this latter dimension of assessment classified him as late-conventional or early-postconventional in terms of Kohlberg's stage-based theory of moral development.

intact memory and attentional capacities, it seems all the more remarkable (and puzzling) that Elliot was not able to, as it were, string all these abilities together in successful real-world deliberations and decisions.

A revealing quote from Elliot points us towards the peculiar deliberative space he came to occupy after his operation. Damasio relates how, at the end of a particular assessment session in which Elliot had generated numerous, perfectly reasonable and sensible options for action, he smiled and said, “And after all this, I still wouldn’t know what to do!” (Damasio, 1994, p49). Clearly, for Elliot, the experience of deliberating and choosing had become something almost entirely open-ended, with little prospect of steady progress towards a smooth resolution and choice. Endless generation and examination of alternatives was one possibility, or else rash impulsive termination of deliberation; and yet none of this because logic and intellect were not on tap to guide deliberation.

Of course, this is not to imply that Elliot’s performance in generating alternatives and anticipating consequences would have looked normal or superior outside of the assessment laboratory. As Damasio (1994) highlights, what Elliot demonstrated himself capable of was working with an initial set of considerations and constraints to generate alternatives for action, projections for likely outcomes and consequences, etc., whereas real-world deliberative and choice conditions present a much more dynamic and complex challenge to the decision maker:

> If it had been “real life,” for every option Elliot offered in a given situation there would have been a response from the other side, which would have changed the situation and required an additional set of options from Elliot, which would have led to yet another response, and in turn to another set of options required from him, and so on… [The] ongoing, open-ended, uncertain evolution of real-life situations [is] missing from the laboratory tasks [on which Elliot performed so well]. (Damasio, 1994, pp49-50)

This is a very different open-endedness to that experienced by Elliot from within his deliberative stance on the world with no obvious route out. Instead, this is the natural open-endedness of choice situations in (especially social) real-world contexts in which the efficient deliberator needs to efficiently navigate the dynamic landscape of options and consequences.

As related earlier, Damasio’s suspicion was that cold-bloodedness, Elliot’s apparent lack of emotion and emotional responsiveness, might hold the key to unravelling the mystery behind

the unravelling of this poor man's life:

> I began to think that the cold-bloodedness of Elliot's reasoning prevented him from assigning different values to different options, and made his decision-making landscape hopelessly flat. It might also be that the same cold-bloodedness made his mental landscape too shifty and unsustained for the time required to make response selections, in other words, a subtle rather than basic defect in working memory which might alter the remainder of the reasoning process required for a decision to emerge. (Damasio, 1994, p51)

Much of the balance of *Descartes' Error* is spent developing this idea, culminating in Damasio framing his *somatic-marker hypothesis* as an explanation of both the decision-making defect in cases like Elliot's, as well as an account of the normal, healthy role of emotion in ordinary decision-making[217]. The essence of this hypothesis is that, according to Damasio (1994), ordinary decision-making is enabled, guided and streamlined by options presenting themselves to the deliberator as emotionally tagged, flagged or marked. (The somatic dimension of this tagging or marking derives from Damasio's strongly body-based account of emotion.) These markers will involve positive or negative affect, and they serve as a kind of flag for approach or avoidance of a given choice option. Healthy deliberators and decision-makers use these markers to then efficiently navigate their decision-making landscape, avoiding those options that have been negatively flagged while whittling down the remaining possibilities amongst the positively or more neutrally marked options.

Damasio (1994) is not always as generous with his unpacking of his theory as he is with his case material, but we can get a reasonably clear picture of the basic emotion-reason mechanism he envisages. As options come to mind in deliberation, they tend to come with somatic, emotion-laden markers attached. These would, in the first instance, be learnt associations and evaluations based on past experience, including past deliberations on consequences of different options. So, if lying to a colleague to escape censure at work has been thought through as a (bad) option in previous choice situations, it might present itself as an option now, but with a negative emotional tag. The tag obviates the need to repeat the cognitively demanding process of thinking that option through in detail. Similarly for options that have positive emotional tags based on prior action, experience and/or deliberation. In addition to remembered associations and valuations, somatic/ emotional marking will take effect within current deliberations, such that when a desirable or undesirable consequence is encountered in the process of thinking an option through, that option will be appropriately and efficiently tagged in all ensuing deliberations. Not only is the deliberative landscape thus

[217] See also Bechara, Damasio, Damasio and Anderson (1994), Bechara, Damasio and Damasio (2000), and Bechara and Damasio (2005).

given emotional shape and, as it were, a gravitational force of approach and avoidance, but it is further rendered highly efficient (most obviously in terms of demands on working memory), especially in just the kinds of interactive, looped, dynamic real-world deliberative scenarios Damasio described earlier.

In summary, Damasio (1994) offers us a compelling, empirically-based theoretical argument towards both a positive and a negative thesis. The negative thesis, that is the origin of the title of his book, is that the importance of affect (which he understands in terms of 'somatic marking') in real, effective thinking and decision-making helps expose one of Descartes' fundamental errors:

> …the abyssal separation between body and mind, between the sizable, dimensioned, mechanically operated, infinitely divisible body stuff, on the one hand, and the unsizable, undimensioned, un-pushpullable, nondivisible mind stuff; the suggestion that reasoning, and moral judgement, and the suffering that comes from physical pain or emotional upheaval might exist separately from the body. Specifically: the separation of the most refined operations of the mind from the structure and operation of a biological organism. (Damasio, 1994, pp.249-50)

At the same time, Damasio (1994) sees the evidence he presents about the link between emotion, feeling and reason as providing significant support for his overarching positive claim:

> …that the comprehensive understanding of the human mind requires an organismic perspective; that not only must the mind move from a nonphysical cogitum to the realm of biological tissue, but it must also be related to a whole organism possessed of integrated body proper and brain (*sic*) and fully interactive with a physical and social environment. (Damasio, 1994, p252)

Taken together with the details of his case material and his somatic-marker hypothesis, Damasio's work points to an important claim about feeling and reason that we need to accommodate in any account of conscious RD agency – namely that an over-intellectualisation of reason, deliberation and choice, especially in so far as reason becomes or is recommended to be divorced from emotion, will leave us with a distorted, disembodied conception of agency that is detached from the biological realities (and complexities) of embodied human agency.

Blackburn (1998) is interested in precisely these kinds of implications of Damasio's work for our understanding of the role of emotion in reasoning and decision-making. For Blackburn:

> …the most fascinating result of Damasio's work… is the extent to which 'higher-order' decision-making has to harness the limbic system [implicated in primary emotional response] in order to work at all. The clinical evidence is that when the primitive system [associated with the limbic system] is disrupted, then the higher-order decision-making system collapses with it. Unless the outcomes of action that the cognitive system can identify come 'somatically marked' [i.e.

> emotionally tagged], the decision-making system malfunctions. The whole point is that there is no dualism, in which the one floats free of the other. (Blackburn, 1998, p.129)

Thus, Blackburn sees in the clinical data as providing empirical support for the claim that emotion plays an inextricably crucial role in decision-making, giving shape and contour to the decision-making landscape, and biasing what we notice, think about, remember and deliberate over when we are confronted with choices between courses of action. To put the point more negatively, the data suggest it is a mistake to attempt or assume a divorce between reason and emotion, or to assume or claim that reason can (or ought to) magically detach itself from emotional systems (primitive or otherwise) so as to entirely transcend the influence of body-based emotion, because it is not heightened or improved or 'ultimate' rational control that would follow such disconnection, but severe malfunction. In Blackburn's (1998, p.131) words, "Without emotions, the will is rudderless."

This is the critical lesson of Damasio's (1994) work. There is a strong tradition in Western thought that would have reason lording it over more 'primitive' motivational and evaluative systems such as the emotions, and recommending a dispassionate deliberative stance in which emotion might be recognised (as if from the detached position of an observer) and even accorded some weight in a rational calculus, but only once 'translated' into the cognitive currency of reasons. Damasio's work suggests that what this tradition recommends by way of detachment and distance would amount to pathology – a life that, like Elliot's, is variously characterised by poor decisions, rash decisions, apathetic inertia, and interminable deliberations, in a decision-making landscape that has become flat and featureless. This is not to recommend giving unrestricted sway to emotion, or to prescribe 'hot' responses to life's challenges situations based purely on emotional reactions, pushes and pulls. Damasio is, after all, offering us a theory of normal, healthy deliberation and choice, and thus a picture of reflective and deliberative agency that includes thought, reasoning, and an adaptive weighing up of options and considerations. His data and arguments suggest, however, that emotion must be given its proper place at the heart of these lived, embodied, dynamic processes, and not be consigned to some corner of a balance sheet in a detached cognitive calculus[218].

[218] My favourable view of Damasio's theory, and the significance I see it having in the context of my project, do not imply uncritical acceptance on my part. Damasio's is an empirical hypothesis, and it must ultimately answer to empirical data and testing. For a comprehensive critical review, see Dunn, Dalgleish and Lawrence (2006).

Realistic Constraints Without Scepticism

Simon Blackburn asserts, bluntly, that "philosophers are professionally wedded to the power of intelligent thought" (Blackburn, 1998, pp.261). He might well have framed this sentiment in terms of an ardent attachment to the significance and power of reflection. All too often, we assume, assert or recommend a reflective stance that turns us away from an active, dynamic engagement with the world (including our peers) in favour of an inwardly-directed introspective gaze at (what we think are) the contents of our internal worlds. In the process, we objectify our mental states, making as though these were entities or events in us, forgetting the corrective offered to us by Strawson and Blackburn (see Chapter 7) that these are instead states that we occupy, and that shape our interaction with and relation to the world.

The present chapter has offered a number of reasons for being cautious in the assumptions we make about the more obviously cognitive, interpretive powers of reflective consciousness, and the extent to which we might try and disengage these from either the world or dynamic embodied engagements with it. Coming at these matters from a more philosophical angle, Blackburn reinforces these cautions by addressing what he thinks if the fundamental mistake about the nature of deliberation and the deliberative stance:

> Typically, in deliberation what I do pay attention to are the relevant *features* of the external world… I don't also pay attention to my own desires… My own concerns and dispositions determine which features [of the world, of the situation] I notice and how I react to them… There is not typically a *second-order* process of standing back, noticing that the cost is obsessing me, and deciding to endorse that fact about myself, or alternatively deciding to try to change it. (Blackburn, 1998, pp.253-4)

Note that Blackburn does not deny the possibility or potential value of introspection, or of various second-order processes directed at myself, my interests and concerns. What he asserts are claims about the typical case, about the normal deliberative position taken up by the conscious, reflective agent – a position or stance that engages the world, where reflection is on features of the world, the agent's reactions to it, and on possibilities for action to further the agent's interests and concerns.

Finally, Blackburn (1998) recognises two important limits to the value of gathering and reflecting on information for purposes of deliberation and decision, even while recognising that there is a quite general value to be attached to informed, reflective decision-making:

> [First] of all I recognise that there are many occasions when spontaneity is more important than laborious collection of facts and reflection on outcomes. And secondly, there are occasions where

> too much information itself distorts our judgement: even the fairest skin looks lumpy and blotchy from too close up, a close acquaintance with the processes of mastication and digestion can destroy the proper enjoyment of food, a fond and vivid representation of how pleasant it is to be rich may make me susceptible to bribery, and so on. (Blackburn, 1998, pp.261-2)

There are perils to thinking too much and knowing too much. The value of conscious reflective and deliberative agency, conceived along strongly cognitive, articulated and reason-based lines, is this restricted and qualified. Are there, perhaps, other domains or spheres of reflective consciousness that have a less qualified and restricted relevance to a distinctive form of human agency, and to significant degrees of freedom in action? That will be the topic of Chapter 10.

Chapter 10

Realistic Self-governance II: Freedom, Imagination, and the Externalised Mind

A central thrust of the previous chapter was to raise a number of concerns and cautions in how we characterise the reflective and deliberative form of conscious agency that is distinctive of human beings, and that acts as an important grounding for claims that we are free agents. These cautions, primarily motivated by empirical evidence, but backed up by compelling philosophical considerations, jointly suggest that we should avoid the tendency to over-intellectualise human agents, both because evidence suggests they (we) may come up short of expectations, but also because we may find that an overly intellectualised, reflective and detached deliberative stance on our lives and the world is not normatively desirable either.

At no point has it been suggested that conscious reflection and deliberation of a more intellectualised and cognitive variety is *unimportant* to distinctively human agency, or to claims that we are free agents. This would be absurd. There is a vast literature on reason-based deliberative agency that makes valuable contributions to our understanding of various distinctive features of human agency, and to how we might claim various important degrees of freedom in our interactions with the world on the basis of these features, so long as we keep certain cautionary notes (such as those sounded in Chapter 9) in mind, and thus avoid committing ourselves to claims that we are free by virtue of our attaining some form of hyper-reflective, hyper-rational agency (HHA).

The aim of the current chapter is thus not to revisit this literature or these issues. Instead, as advertised in the first parts of Chapter 9, I want to propose and argue for the importance of two less-well explored avenues in which we might find or locate distinctive degrees of freedom besides the more obviously reason-based deliberative aspects of human agency. These avenues are the imagination, and certain externalised aspects of mind. My goal is not to sketch a comprehensive outline for an account of human agency that can secure claims of freedom; nor can I pretend to offer fully worked-out accounts of either imagination or of externalised dimensions of mind. The goal is to explore and test the promise of these

phenomena for the potential contribution they might make to the reframed debate on free agency proposed and outlined in Chapter 5, while at the same time strengthening the responses offered so far to the sceptical arguments and evidence of Chapter 6. Further, both avenues of exploration offer means to highlight the importance of consciousness and occurent mentation to free agency that are often lacking in more reason-based deliberative accounts, thereby further bolstering resistance to the sceptical claim that we are Agent Automatons. I conclude the chapter by delivering on an earlier promise (in Chapter 5) to expand on my speculative hypothesis as to how we might link imagination and indeterminism within a future incompatibilist defence of free will.

The Neglect of Imagination

The Harvard University Press publication blurb for Colin McGinn's (2004) book on imagination, *Mindsight*, suggests that the topic of imagination has a rich history of exploration in philosophy until the rise of contemporary analytic philosophy of mind. Whatever the truth of this claim for the philosophy of mind in general, it certainly has some warrant as an empirical claim about the prominence – or, rather, the neglect – of imagination in discussions of free will and agency.

An indication of the degree to which imagination forms a neglected topic within (especially philosophical) discussions of free will and free agency can be found through rough-and-ready surveys of some significant texts published in the last decade[219]. Robert Kane has edited two important volumes on free will – *The Oxford Handbook of Free Will* (Kane, 2002c) and *Free Will* (Kane, 2002b) for Blackwell. Neither of these texts has an entry for 'imagination' in their index. An electronic search of the Blackwell volume (via Amazon UK) revealed not a single reference to imagination in the 310 pages of Kane (2002b); while a similar electronic search of the 638 page OUP volume revealed just two uses of the term: one in a footnote to Russell (2002), in which imagination is listed as one of the 'natural abilities' discussed by Hume, and the other in Ted Honderich's (2002) contribution to this volume, where "abstract concepts of the imagination" (Honderich, 2002, p463) is one of a long list of speculative ideas discussed by physicists with regard to the nature of quantum events. In 948 pages of significant historical (Kane, 2002b) and contemporary (Kane 2002b, 2002c) discussions of free will, not one contributor saw fit to discuss imagination (certainly not by name) as an

[219] At a bit of a stretch, we could consider the survey that follows to be in the spirit of the quantitative approach to the history of philosophy experimented with by Nichols (2007).

important capacity or faculty within human psychology that might have relevance to how we understand, account for and defend free will.

The index to *Freedom Evolves* (Dennett, 2003) lists just two entries for imagination. On page 179, in the midst of a discussion of Dawkin's idea of a meme, Dennett claims that access to memes opens up "a world of imagination to human beings that would otherwise be closed off" (Dennett, 2003, p179). More promisingly, Dennett's other explicit reference to imagination occurs in the context of a discussion of what he claims is our distinctive capacity as humans to appreciate and exploit aspects of and opportunities for design, instead of merely being subject to the blind 'research & development' of natural selection:

> We are the only species whose members can *imagine* the adaptive landscape of possibilities beyond the physical landscape, who can "see" across the valleys to other conceivable peaks. The mere fact that we're doing what we're doing – trying to figure out whether our ethical aspirations have any sound anchoring in the world science is uncovering for us – shows how different we are from all other species. (Dennett, 2003, p267, italics in original)

So Dennett is happy to acknowledge a role for imagination in conceiving of better worlds, and better arrangements for society. These are important and welcome ideas, but in the context of both my project and Dennett's project in *Freedom Evolves*, it represents a rather late and relatively sophisticated application of imagination to issues of freedom and agency[220].

Psychologists don't pay much attention to imagination either when they turn their attention to questions of free will. In a recent collection of papers on the topic of psychology and free will (Baer, Kaufman & Baumeister, 2008), 'imagination' is once again conspicuous in its absence from the volume's index, and an electronic search yields only three hits. Two of these are occurrences within reference lists (i.e. titles of cited works), and the third occurs in a quotation used by Mele (2008) where the relevant sense of 'imagination' involves reference to the alleged impact of Libet and Wegner's studies on the philosophical and scientific imagination. Once again, then, this evidence suggests that imagination is not deemed an especially important aspect of human psychology to explore in the context of discussions of free will.

[220] Dennett makes numerous references to and uses of 'imagination' throughout *Freedom Evolves*, as an electronic search like the ones carried out for Kane (2002b) and Kane (2002c) will reveal. However, most of these references to imagination occur in contexts where Dennett is talking about it as a philosophical tool. For example, he describes his uses of fables and examples in books like *Brainstorms* and *Consciousness Explained* as "exercises in imagination-shifting" (Dennett, 2003, p223), exercises that Dennett thinks many of his readers are reluctant to engage in.

These surveys are both superficial and without any claim to statistical representivity. They are suggestive, though, of a robust trend. This trend is confirmed by a far more wide-reaching and representative online keyword search of *The Philosopher's Index* (via the EBSCO portal) combining "free will" and "imagination" as search terms. The search yielded just three hits: two with a historical and interpretive focus (one on Spinoza, one on Schiller), and the third being a work on symbolism in fiction.

Of course, too much emphasis on these 'results' could be thought of as something of a double-edged sword. It might be objected, for example, that what these keyword searches of indexes, texts and databases demonstrate is not an area of neglected theoretical and philosophical exploration ripe for development and application to questions about freedom, but rather a firmly established consensus that imagination just is not central to issues of free will and agency. Is there something more that can be said, *prima facie*, in favour of the idea that imagination ought to play a more prominent role in our attempts to characterise and defend claims of free agency?

Speaking up for Imagination

I want to cite three initial sources of support for taking a sustained look at what imagination might contribute to a richer conception of free agency. The first has already been cited above – that is, Dennett's (2003) idea that imagination is crucial to our ability to conceive of alternative 'landscapes' for agency, society, and the 'design' of both. Dennett's reference to imagination in this context is all too brief, and is situated in the midst of his strongly evolutionary account of the emergence of freedom, with relatedly strong emphases on engineering and design perspectives in the context of various evolutionary 'arms races', whose details will take us too far away from our current concerns[221]. But his reference to imagination at least makes an explicit claim as to the importance of imaginative capacity and imaginative activity to our being able to conceive of alternative possibilities for ourselves – a claim that seems surprisingly absent from the traditional debate on free will.

A second source of support comes from John McCrone's contribution to Libet, Freeman and Sutherland's (1999) volume *The Volitional Brain*. In the conclusion to his paper, McCrone

[221] We will visit certain aspects of Dennett's account in greater detail below, as a foil for discussing the significance of Donald's (2001) notion of kinematic imagination.

(1999) highlights the capacity of brains to appropriately focus attention and draw on stored knowledge to deal with the demands of the moment, while noting that different species are not equal in this regard:

> [The] difference between animals and humans is that we can make our brains react to imaginary contexts – we can do things like think about what it would be like to go into a room and find a handbag left on a chair. And we also habitually bring a socially-expanded sense of context to bear on each moment – our social training will make us imagine what people would say if they saw us rifling through that handbag, even if we just happened to be sneaking a peak out of curiosity… [Humans] have created an extra, socialized, level of filtering that all brain planning must pass through before the possible becomes translated into the actual. (McCrone, 1999, p257)

McCrone's claims about imagination point to the possibility of a significant shift in emphasis from the obviously cognitive, reason-based perspectives on reflective deliberation presented at the beginning of Chapter 9 (Korsgaard, Compatibilism, etc.). Instead of responsiveness to *reasons* being at the core of a distinctively reflective deliberative human agency, it is the activity of and responsiveness to an active and productive imagination that is critical. Imaginative agents can, of course, also be appreciative of and responsive to reasons. They can be reasonable, and more or less rational. But they are first, in a significant sense, imaginative agents capable of imagining and responding to not just the actual, the present and the remembered, but also to the possible, the absent and, perhaps, even the impossible. Relatedly, McCrone's claims suggest that what we might call the ethical point of view is primarily an achievement of imagination – imagining the point of view and the responses of others to our actions. Supplemented with, for example, the perspective on emotional response and marking offered by Damasio (1994) (see my Chapter 9), such that the *imagined* scene of being found rifling through the handbag provokes an *actual* (if fleeting) emotional response of shame, this suggests a reflective process, even what we could call a deliberative process, whose psychology or occurent reality is *not* a cognitive one of recognising and articulating reasons for action before deliberating over and acting on these *qua* reasons.

My third source of support is Colin McGinn (2004). In *Mindsight*, he notes (all too briefly) what he sees as the crucial connection between imagination and free will in the midst of a discussion of the role of imagination in meaning:

> Chomsky long ago made the point that linguistic use is "stimulus free" – that is, it is not elicited by some ambient stimulus in the manner of a conditioned reflex. We can clearly speak of distant and absent objects, of the non-existent, the past, the future, and so on. An utterance is not an automatic response to some environmental contingency. In this respect it is nothing like a percept, which is emphatically *not* stimulus free. But of course stimulus freedom is what the imagination specializes in: absence, non-existence, revision, outright invention. We might even say that the whole point of the imagination is to escape the domination of the stimulus – to transcend the present (temporally and spatially). Imagination frees up the mind to escape what is impinging on

> it. It is, indeed, part of what makes us free agents. So the stimulus freedom characteristic of language is nicely captured by introducing imagination into the heart of understanding. (McGinn, 2004, p153, italics in original)

McGinn does add a note to this frustratingly brief assertion about imagination being part of what makes us free agents. In this note, he expands on his claim as follows:

> Imagination is what presents the mind with alternative courses of action – the envisaging of possible futures. There is some merit in the idea, favored by the Romantics, that the imagination is the *primary* locus of human freedom: it is what makes our overt actions free by offering us alternatives, and it is itself an instance of free action, as we use it spontaneously to create all manner of marvelous mental products (literature, music, science, philosophy, etc.). Certainly, imagination is the most weightless and unconstrained of human faculties, the most fleet and feathery. (McGinn, 2004, p195, italics in original)

McGinn's comments reinforce the points made earlier in relation to Dennett's comments on imagination – that is, the emphasis on imagination in understanding our capacity to conceive alternative possibilities, alternative futures – while also highlighting the apparent 'weightlessness' and freedom of imagination that we might traditionally associate with its more artistic employment, but that McGinn is also associating with not only other intellectual exploits, but with a kind of creative freedom in agency *per se*.

Together, these three authors point towards a number of ways in which imagination ought to have achieved greater prominence in our understanding of freedom and agency. It will not be possible to expand on and explore all of these ideas in what remains; at any rate, my suggestions and proposals in this chapter must remain somewhat programmatic, pointing to avenues for future work. My immediate interest in running with the idea that imagination is central to a form of agency that is distinctively human, as well as crucial to an understanding of our freedom, is in order to reengage with a number of issues from earlier chapters: questions about the nature of bodily control, about self-reflective consciousness, and questions about the intellectualisation of agency through an (over-)emphasis on reflective articulation of, and deliberation over, reasons. I will first revisit these issues by way of a less obvious form of imagination – imagining the body in action, or what Merlin Donald (2001) has called *kinematic* imagination. Thereafter, we will return to some more familiar senses of imaginative activity in examining empirical work on imagination and its significance in child development.

Consciousness, Mimesis and Kinematic Imagination

As briefly discussed towards the beginning of the previous chapter, Dennett (2003) links the emergence of new capacities for self-monitoring – capacities that will be crucial to our freedom – to the evolutionary emergence of communication:

> It is only once a creature begins to develop the activity of communication, and in particular the communication of its actions and plans, that it has to have some capacity for monitoring not just the results of its actions, but of its prior evaluations and formation of intentions as well. (Dennett, 2003, p.248)

This emphasis on communication makes Dennett's (2003) theory heavily dependent on language and linguistically-mediated thought and communication. It also enables him to make the claim that the self of agency is a kind of cultural and linguistic invention that provides a useful 'user interface' in the ensuing social interactions of giving and demanding reasons. This is an outside-in account of the emergence of distinctively human forms of self-monitoring and control that would (i) locate our most distinctive degree of freedom in agency only after the emergence of language, and (ii) do so in a way that renders the 'self' of the associated self-monitoring and control a kind of user-illusion.

It is my contention that both of these implications can be avoided (or, in the case of (i), at least seriously re-evaluated) by exchanging Dennett's (1991, 2003) particular brand of evolutionary theorising about agency from within a sceptical stance on consciousness with an alternative account of the evolution of human consciousness developed by the psychologist Merlin Donald (2001) that has, as I see it, important implications for our understanding of agency and imagination. Donald's theory is too extensive to summarise here in any detail. Its primary significance in this current chapter derives from his proposed notion of kinematic imagination and its role in the emergence of a novel and distinctive combination of mind and culture that he calls mimetic culture.

Donald's (2001) account of kinematic imagination forms part of his more general theory of the evolution of human consciousness, a theory that attempts to trace a series of cognitive transitions importantly linked to the emergence of what Donald calls 'cognitive communities' and shared knowledge networks that are the stuff of culture. Donald ultimately sees symbolic thought and language as *by-products* of a series of prior historical developments in hominid evolution that progressively established patterns of bonding and cooperation through shared

cognitive activity, and the establishing of shared, distributed knowledge networks:

> The scenario of human evolution seems to be one of tension between culture and conscious capacity, with culture steadily pushing that capacity to the edge, so that it continuously expanded. Culture was a radically new presence, and the mind kept adjusting itself to the new reality of distributed cognition [within cognitive communities]. The result of that tension, in the long run, was the emergence of a symbolizing mentality. (Donald, 2001, p.260)

In other words, Donald offers a theory that charts the emergence of ever more distinctively human forms of conscious capacity, where the demands of shared and distributed patterns of cognition within cognitive communities provide the most crucial drivers for change.

Donald's (2001) theory posits three historical transitions in the evolution of hominid and human consciousness and culture. The first transition involved the emergence of mimetic capacity and the establishing of what Donald calls the mimetic framework of human culture. The second transition, involving the emergence of language and symbolic representation, is associated with the emergence of oral and mythic culture; while the third transition involved the invention of (external) symbolic technologies and the associated emergence of theoretic culture. The latter stage is held by Donald to date at least from the time of Ancient Greece, and is the stage in which contemporary humans still operate. (I discuss aspects of this stage, and specifically symbolic technologies, towards the end of the current chapter.)

Our primary focus is on what Donald (2001) hypothesizes as being the first transition. The core of the first transition corresponds to the emergence of a novel kind of cognitive capacity involving the extension of conscious executive control into the domain of action – mimetic skill. In terms of the fossil record, Donald (2001, p.261) associates this intellectual transition with "a general increase in brain size"; and, in terms of archaeological finds, with the emergence of stone tools, a meat-heavy diet (including larger game), wooden spears, toolmaking sites, seasonal hunting camps, and continuously occupied fire sites. According to Donald (2001), these provide evidence not only of a communal style of living, but also of *skilled cooperation* and *shared knowledge systems* that have, at best, only faint echoes in our contemporary primate cousins (I am thinking here of evidence of simple tool use in the breaking of nuts, as well as coordinated hunting activities, amongst chimpanzees).

The emerging 'culture' Donald has in mind is one of public action, underpinned by mimetic expressive skills, but without language or symbols. To understand this emergence of a

distinctively hominid/ human culture, we need to look more closely at what he means by 'mimesis':

> Mimesis is the result of evolving better conscious control over action. In its purest form, it is epitomised by four uniquely human abilities: mime, imitation, skill, and gesture. These are direct offshoots of the expansion of the human executive brain system… (Donald, 2001, p.263)

Donald (2001, p. 263) defines mime as "the imaginative re-enactment of an event", with central examples of mime being the pretend play of children, as well as the re-enactment of emotional events. While these are not examples of fully symbolic mental representation, they are nevertheless representational in the sense that they implicitly refer to something other than the performance itself – namely previously witnessed or experienced events. These are reproduced, by way of role playing, through a combination of memory of the original event and cognitive control over action (including emotional expression).

Donald (2001) sees the second aspect of mimetic capacity – imitation – as more complex and demanding than simple mime because it involves the replication of events that have some instrumental purpose. While mime involves the mere copying of an action or action sequence, imitation requires some recognition and understanding of the purpose or objective that another person had in mind when performing the action. Thus, in Donald's (2001, p.264) chosen example, a child may be able to mime the act of sharpening a wooden spear long before they are able to imitate this action because it takes time for the child to grasp the purpose of the activity.

Closely related to mime and imitation is the third aspect of mimesis – skill. Donald (2001, p.264) sees skill as resulting from "rehearsal, systematic improvement, and the chaining of mimetic acts into hierarchies… [When we learn a skilled craft or activity], we must learn a set of basic action sequences, generalise them, and rehearse them until they become second nature." This process draws on mime and imitation, but also extends beyond these more simple processes. That is, while we may begin our learning of a skill by imitating its production by another, and rehearse the activity by miming our own previous performances, we may find that success in our own case requires some modification of the action sequence we have observed in others, or the inclusion of novel ways of approaching the task.

Finally, Donald (2001) sees gesture as something that stems naturally from the combination of mime, imitation and skill. What marks out gesture as distinct, especially in comparison to

mime, is that gesture is usually an explicitly communicative, intentional act. We can think of gesture as working through the shared use of mime to iconically represent objects, people or events through mimetic re-enactment. Thus, in Donald's (2001) example that he borrows from developmental psychologist Linda Acredolo, a child can learn to represent a fish by miming a stereotyped puckered mouth. As gesture becomes more developed with age, more complex representations can be constructed using sequences of (iconic) gestures. Donald (2001) does not view the communicative acts involved in and enabled by gesture as truly symbolic or linguistic. Nevertheless, gesture takes us into a mimetic realm that is more clearly communicative, and where possibilities for conventional representation of persons, objects and events through shared mimetic practice can form the basis for not only transmission of skills and knowledge within a community, but also a basis for shared cultural practices and traditions.

With these details in mind, Donald proposes mimesis as the first distinctively hominid transition towards forms of expression, coordination, knowledge sharing, and systems of convention and tradition within communal cognitive networks. Ultimately, Donald views the emergence of mimesis and mimetic culture as a move or transition that, over time, helped bring the emergence of language within the 'zone of proximal evolution'[222] for hominids. But the significance of mimetic capacity should also be noted at the level of brain and cognitive function:

> Mimetic capacity was primarily the result of merging the executive brain with the action brain, when the hominid executive brain system extended its anatomical territory into the frontal and subcortical regions that control voluntary action. This anatomical change had major social ramifications. Active social networking, even of the exclusively mimetic variety, makes heavy demands on attention and memory, and we can corroborate this with evidence drawn from studies of child development… Children become excellent mime artists and actors, long before they can verbally describe or reflect on what they are doing. (Donald, 2001, p.266)

In other words, Donald sees mimesis as strongly associated with the 'leveraged takeover' described by Deacon (1997) (as discussed in Chapter 8) of cortical and subcortical structures in the brain controlling action through mechanisms such as displacement, whereby the general increase in the relative size of frontal areas in hominid and human brains has the ontogenetic effect of bringing formally autonomous (or largely autonomous) behavioural control systems under the influence of cortical and, especially, frontal processing, thus allowing for greater executive control of a wider range of bodily activity. Donald suggests

[222] An evolutionary application, or analogue, of Vygotsky's developmental notion of a zone of proximal development.

that this shift to greater executive control is echoed in the developmental histories of contemporary children. That is, the level of executive control of action required for the mimetic activities of children – imaginary or fantasy play, role playing, etc. – remains an important achievement as children enter the social world of custom, convention and role-taking that are the precursors of language acquisition and truly symbolic thought.

Mimetic activity in children also points to fundamental ways in which human consciousness is able to shift from an episodic focus on present experience to the representation of past and future. Fantasy play involves, amongst other things, the mental and behavioural construction of scenarios that, while grounded in the child's experience, are often generated and sustained in the absence of salient environmental input or prompting. The child enters into the world of the conditional, the as-if world in which scenarios can be played out in mind and behaviour, fostering not only the performance and rehearsal of actions, but also the capacity to anticipate alternative scenarios (instrumental, affective, or social) that might be contingent on different courses of action.

The characteristics of mimesis highlighted by Donald (2001) have a familiar ring to them. What Donald is describing in historical-evolutionary terms is the emergence of a distinctively human consciousness that can be thought of as self-consciousness (Searle, 1997), higher-order consciousness (Edelman, 2004; Edelman & Tononi, 2000) or extended consciousness[223] (Damasio, 1999). It is consciousness that involves an explicit sense of self (what Damasio ,1999, helpfully calls the autobiographical self) and the capacity to construct and maintain in consciousness scenes of past or future scenarios (Edelman & Tononi, 2000). And, while most theorists allow that some aspects of this consciousness are likely to be shared by at least some of our mammalian cousins (especially chimpanzees), full-fledged extended consciousness is only to be found in humans where it is accompanied by the representational and semantic abilities we associate most closely with language.

And yet, where many theorists see advanced semantic and specifically linguistic abilities as the clearest marker of a distinctively human extended consciousness, Donald (2001) thinks that it is our pre-linguistic capacity for mimesis that represents the most distinctive break

[223] I will use Damasio's term 'extended consciousness' wherever possible, since 'self-consciousness' has the effect of focusing too much on the self-related aspects of this form of consciousness, and 'higher-order consciousness' might be confused (in philosophical circles) with an attachment to a higher-order thought (HOT) theory of consciousness.

between humans and our evolutionary forebears and cousins. For example, while Donald (2001) is sympathetic to much that Deacon (1997) has to contribute to our understanding of the evolution of mind, brain and language, he thinks that Deacon's own account of the emergence of language is flawed because it focuses too much on issues of communication, representation and reference, and posits a transition directly from indexical modes of reference (characteristic of, for example, many animal alarm calls) to the emergence of full-fledged symbolic reference, thereby ignoring the significance of pre-linguistic and action-centred mimetic capacities and associated aspects of a mimetic culture[224].

The significance of such differences cannot be overemphasised. For Deacon (1997), the most important pressures shaping the evolutionary emergence of language relate quite directly to issues of reference, representation and communication. Selective pressure to communicate more effectively using fledgling symbolic systems drives selective pressure for more cortical influence and control over the activities of the tongue, larynx, and the (otherwise automatic) process of breathing. For Donald (2001), the cortical influence and control of action required by mimesis precedes the emergence of symbolic communication. Distinctively human consciousness is first a consciousness of action, and the agent-self in action, before it is a consciousness enhanced by symbolic representation and reference.

The exercise of (pre-linguistic) kinematic imagination is a relatively straightforward process that should strike us as familiar. It involves the capacity to recall or imagine a performance of the body (one's own or that of a model to be mimed or imitated), to attempt the same performance, and then recall and review the latter performance in comparison to the original recalled or imagined performance. As Donald (2001) points out, this is a sequence that we generally employ in our learning of just about any skilled activity, from diving into a pool to learning to play a musical instrument, or a new musical piece, or to drive a car.

[224] In a note (Donald, 2001, p341) on the co-evolution of brain and language, Donald comments: Deacon (1997) proposed a theory of language evolution that tries to finesse the need for a mimetic preadaptation. He argues that hominids made the transition from primate "indexical" representation (as shown in operant conditioning) to fully symbolic representation (as shown in language) in one continuous evolutionary progression. This idea does not account for the metaphoric nature of language, the existence of mimetic skill, or the fact that language normally develops in a mimetic framework, and operates by metaphoric or mimetic principles. Moreover, it cannot scale the wall set up by the primate zone of proximal evolution.

Described in this way, kinematic imagination might sound so ordinary as to seem mundane, but its emergence has great significance for the economies of mind and agency. Donald (2001) claims that the anatomical changes described by Deacon (1997) brought with them a shift in the terrain and focus of conscious processing, from perception and short-term memory towards the conscious control of behaviour:

> Whereas primate consciousness had occupied a largely perceptual domain and was directed mostly at the outside world, hominid conscious capacity invaded the domain of action. The executive brain system improved its ability to monitor the state of the physical self and gained access to much more detail, so that precise attention could be paid to the body's own movement patterns. (Donald, 2001, p.270)

New modes of cortical modulation and influence over bodily activity became anatomically and cognitively possible; and attention and memory resources were now required for the monitoring and management of cognitive and bodily activity over the timescale of repeated action sequences, instead of only the shorter-term demands of conscious perception.

It is not only attention that was redirected away from the outside world towards an agent-centred world of action. The exercise of kinematic imagination also represents a significant shift in emphasis when it comes to the dynamics and focus of learning, and this shift is also one from the world outside to a more internal, mental world. In the place of the usual extrinsic focus on the reward or punishment that follows an action, the primary focus in learning through the use of kinematic imagination is on the *form* of the act itself (Donald, 2001), and the extent to which the performance succeeds in approximating the imagined or recalled 'ideal' or template. Refinement of bodily control and skill through the use of kinematic imagination provides the agent with a bootstrapping mechanism for learning that is thus at least somewhat removed from the contingencies of reward and punishment.

The Significance of Kinematic Imagination

Mimesis and kinematic imagination involve a turning of attention towards the body in action, as well as an ability to shape the body's movements in an attempt to approximate a remembered or imagined ideal. As Donald (2001, p270) puts it, "the first rung in our distinctively human ladder of awareness is the physical self, a supraordinate form of body awareness, born of a need to refine action." Donald's theory is, needless to say, empirically defeasible – although it is not clear that it need be any more or less so than the evolutionary account of freedom and agency offered by Dennett in *Freedom Evolves* – and it will need and benefit from greater integration and triangulation with other relevant evidence, including

advances in the neuroscience of action control. For present purposes, however, it will be sufficient to note the alternative perspective that Donald's theory can offer on some of the central issues in the preceding chapters, and particularly on those of this and the previous chapter.

Donald's ideas about the exercise of kinematic imagination have relevance, for example, to those who would challenge our claims of conscious agency on the basis of evidence for pervasive automaticity – recall Bargh and Chartrand's (1999) provocative article title, "The Unbearable Automaticity of Being." While Bargh and Chartrand (1999) tend to focus their attention on alleged cases of what might be called cognitive automaticity, a much more common context in which we talk about automaticity is that of action – especially contexts of learning and performing *skilled* actions. As Donald (2001) describes it, the exercise of kinematic imagination, with its cycles of performance, review and rehearsal, is critical not only to mastery of the relevant bodily repertoires demanded by a given skill, but it is further a normal part of the process whereby more and more aspects of such performance become automated, thus freeing up resources of attention, memory and imagination, so that these are available for the higher-order task of monitoring and tweaking the overall 'shape' and quality of the performance. On this view, automaticity in skilled performance is, under normal circumstances, friend and aid to effective conscious control of voluntary action, not its enemy.

I am a good driver because the vast majority of micro-skills, routines and sub-routines involved in my driving a motor vehicle became largely automatic years ago. I am not such a great drummer, in part, because I lack the experience and practice required to have automatized enough of what is required of my four limbs while drumming to different rhythms. When I am in my car, I have enough attention and memory resources to concentrate on where I am going and on the behaviour of my fellow road users; on my drums, intense concentration on the activity of one limb often comes at the expense of losing the rhythm for a moment, or in its entirety.

This perspective suggests, then, that automaticity generally acts in service or support of conscious processes of monitoring and control – routines and sub-routines are progressively automatized, wherever possible, so that they may be more efficiently activated without the delays associated with normal timescales of intermediate-term conscious governance (recall

Gallagher's example of the different timescales of responses to a snake in the grass), while also draining fewer cognitive resources of attention and memory in the process. There is no principled reason for thinking that the same logic could not apply to 'mental' acts such as judgement and evaluation. Indeed, just such automatisation of evaluative responses is suggested within Damasio's (1994) theory of somatic markers. For Damasio, the emotions that are primarily implicated in the process of somatic marking within deliberation are what he calls the secondary emotions. Secondary emotions represent acquired emotional associations (as opposed to more basic, inbuilt primary emotions), and on Damasio's account, they are presented as emotional responses that typically follow a mental image rather than the perception of an external stimulus. It is the system of secondary emotions that is disrupted, unreliable or absent in patients like Elliot. For Elliot, what is missing from deliberation is the ability for mental images (of options for action) to automatically trigger emotional markers for those options to aid and ease further deliberations. And while such associations are by no means all the result of prior conscious mentation, at least some of them are[225] – such as the markers that flag an option that is best avoided based on previous (conscious) witnessing or imagining its potential consequences. From this perspective, the efficient operation of somatic markers in deliberation represents, in part, an automatisation of evaluative responses that serves as an aid to conscious processes of deliberation and choice. Such automatising does not threaten conscious agency – under normal circumstances, it makes it more efficient and effective.

As suggested in earlier cited views of the link between freedom and imagination, the exercise of kinematic imagination also involves an explicit awareness of *alternative possibilities for action*, without such alternative possibilities for action needing to be propositionally or linguistically represented. The value and efficacy of kinematic imagination in the development and refinement of skills lies in the possibility of comparing a represented ideal of action (whether recalled or constructed in imagination) with a recalled image of a prior performance, and where the agent has an embodied sense of alternative possibilities for performance. In the absence of such a sense of alternative possibilities for performance, the

[225] Harris (2000) emphasises the importance of our tendency to experience emotional reactions to imagined scenarios, whether these be on the basis of honest testimony (i.e. events and experiences related to us by a witness), fiction (e.g. story-telling) or spontaneous imagination (e.g. playing out imagined scenarios through the mechanism of imaginary friends). On such a view, we might think that most somatic markers reflect a memory of an emotional response, however fleeting, to a scenario that has at some point been either consciously experienced or consciously imagined. I discuss Harris (2000) connecting of imagination with processes of somatic-marker response below.

comparison of images in imagination would be cognitively and motivationally impotent in relation to the prospect of refining skill. One could say that this sense of alternative possibilities for performance is a *condition of possibility* for the effective exercise of kinematic imagination. Moreover, the sense of alternative possibilities for performance is, for the most part, *veridical* because it is grounded in an autobiographical memory of past success in refining performances, together with a growing embodied, performance-based understanding[226] of the underlying anatomical reality of increased cortical influence over motor structures.

Recalling the ideas about reflection and deliberation that I cited towards the beginning of Chapter 9, the present discussion suggests a rather different account of the origins or roots of a reflective, normative perspective on action. The refinement and automatisation of skilled performance through the exercise of kinematic imagination offers us a pre-linguistic and non-reason-based model on which we can understand the internalising of an ideal for behaviour that then acts as both normative guide and constraint for future performances, both actual and imagined. On this view, the distinctive psychological space in which a conscious appreciation and pursuit of normative 'projects' first takes hold is not (following Korsgaard) that of a reflective and reflectively distanced self as deliberator/ legislator, nor (following compatibilism) of a self as reflective endorser, nor (following Dennett) of a self as useful or even necessary fiction in an externally imposed communicative social practice of giving and demanding reasons for action. It is, instead, envisaged as the much more pervasive, embodied, action-centred and playful space of mime, imitation, and the pursuit of skilled performance.

Some Objections

An objection may be anticipated at this point, however. Surely, it might be argued, the space of imitation with a view towards mastery of a type of performance is not at all uniquely distinctive of human agency. Lovers of wildlife documentaries will have no doubt seen many delightful scenes of young predators practicing the art of the hunt. What, if any, principled distinction is there to be made between this kind of repeated performance in pursuit of what can only be called a learnt, advanced skill, and the learning of skilled performance that

[226] Not to be mistaken for a distinctively cognitive or conceptual understanding of these anatomical underpinnings, of course.

Donald (2001) wants to associate with the exercise of a distinctive capacity for kinematic imagination?

One could quibble with the facts and the details of this objection, especially as regards the appropriate interpretation of the behaviour of these young predators. What seems clear, however, is that such learning represents an unambiguous case of trial-and-error learning through a combination of observation (usually of mom), instinct, and actual performances. Donald (2001) sees the development of skilled performance through the exercise of kinematic imagination as something that goes beyond the costly enactive demands of actual trial-and-error learning. The development of skill requires an understanding and appreciation of purpose, and it involves the rehearsal, review and revision of both recalled and imagined bodily performances *without the need for*, or at least *prior* to any further actual performances. It is, as we noted earlier, a pursuit of the appropriate *form* of an action, exploiting our capacity for explicit recall and imagination. It might look similar, on the outside, to the repeated performances of young hunters, but that would be in large part because of the privacy of the imaginative acts involved in the human case.

A more threatening objection, however, might be raised on the basis of Donald's (2001) eagerness to refer to "the executive brain", and his attempts to harness Deacon's notion of a leveraged takeover of subcortical structures by cortical and frontal brain areas with a consequent increase in voluntary control over behaviour. Is this not an assumption or assertion about precisely the kind of centralised executive control that I was cautioning against in Chapter 8? Is this not precisely the image of conscious control mechanisms lording it over the rest of the brain that we are supposed to have argued against in order to expose the flawed assumptions of the sceptical cases presented in Chapter 6?

There can be no quick and exhaustive response to this objection. As I noted in Chapter 8, there is a tension between ideas about self-organisation and decentralised and distributed mechanisms of control, on one hand, and ideas about self-governance, where the latter presumes some sort of unity of organisation, influence and control over the behaviour of mind and body. I offered reasons for thinking that the tension is not one that is threatening to a defence of free agency. Nevertheless, a number of further comments and responses are appropriate at this point.

First, the neural aspects of Donald's theory, and his incorporation of Deacon's (1997) notion of a leveraged cortical takeover, need not imply that Donald wishes to superimpose a singular, centralised executive system housed in frontal lobes from which it sends orders as per the specifications for a hard-assembled instructive system of control. The most that needs to be said about increased cortical projections from especially frontal areas to subcortical areas implicated in voluntary motor control is this: how else would we imagine cortical influence and modulation of action could be achieved except, in part, by means of increased neural connectivity? If we assume, as seems reasonable, that the exercise of kinematic imagination and other important aspects of conscious monitoring and modulation of behaviour are primarily realised by way of cortical activity, then we should expect these areas to project to various motor and other areas in the brain in order that there be a suitable degree of interaction between these different action-related processes. This is not sophisticated neuroscience – it is just the simple (non-dualist) idea that functional relatedness requires some form of neural relatedness.

Donald (2001) and Deacon (1997) are positing phylogenetically novel forms of connectivity, and novel forms of influence and control over bodily behaviour. As we saw in Chapter 8, Deacon (1997) claims that our distinctive level of control over the mouth, tongue and larynx, relative to that evidenced by our closest primate cousins, is directly related to differences in cortical-subcortical connectivity between the relevant species. Similarly, Donald (2001) is hypothesising that the reach of kinematic imagination in the development of skilled performance is related to such novel patterns of connectivity. But neither of these claims need amount to more than the idea that novel functional interactions and influence require novel mechanisms of connectivity and interaction.

A second response to the objection is more speculative, and it connects with the issues and puzzles raised in Chapter 8. I think it plausible that Donald's claims about our distinctive capacity for kinematic imagination not only suggest a pre-linguistic, non-reason-based arena of distinctive normative activity, but further suggest a possible means for positing a type of centralised control that is consistent with the claims of Chapter 8 – especially as regards the likely importance of a variety of control systems, including various systems operating on self-organising and decentralised principles, to the functioning of a complex biological system like a human organism – while also reinforcing the criticisms levelled (in Chapters 7 and 8)

against the assumptions made about conscious initiation and control of behaviour in the sceptical arguments of Chapter 6.

Recall the three features that Schroeter (2004) posited as being characteristic of basic executive control: (i) the agent that is best identified as the system wielding basic executive control is the *conscious self*; (ii) human actions of a relevant order of complexity are carried out in accordance with action plans, and because we have these action plans, basic executive control is characterised by the agent (i.e. the conscious self) being in a position of, in general, *knowing what they are doing*; and (iii) action can be *initiated by central commands* – "the motor system can be directly activated by a command issued by the conscious self" (Schroeter, 2004, p645). By now, we can point to certain key errors, or at least mischaracterisations, that render Schroeter's take on executive control unworkable. By contrast, utilising the conceptual machinery introduced from Chapter 7 onwards, and adding in Donald's notion of kinematic imagination, we can reframe a more workable understanding of executive control.

First, Schroeter is wrong to identify the holder or wielder of executive control with the conscious self, as if the conscious self were some entity or sub-system within the agent[227]. Instead, a workable notion of conscious executive control would attribute this control to the conscious agent – the whole agent, consciously engaged in dynamic intentional interaction with the world. Second, Schroeter is essentially correct in claiming that executive control is characterised by the conscious agent knowing what they are doing. This was the point emphasised in Chapter 7 when clarifying what it meant to act with conscious intention.

However, Schroeter's under-described idea of an action plan, when combined with his claims about direct action initiation via central commands, is problematic. It falls foul of the warnings of Chapter 8 regarding systems of control within complex biological systems like the human organism, and it plays into the hands of the sceptics of Chapter 6 in subscribing to a simplistic, naïve model of conscious initiation and control that is vulnerable to the kind of criticism offered by Wegner.

[227] Dennett's (2003) mantra is useful here: 'If you make yourself really small, you can externalize virtually everything.'

Suppose, instead, that we jettison the claim that basic executive control requires or involves direct activation of the motor system by way of commands issuing from the conscious self; and suppose further that we interpret the idea of an action plan as an imagined performance in the agent's kinematic imagination, not as some set of detailed instructions for activating the motor system. Now we have a centralised monitoring system, where feedback consists in comparisons between real-time and recalled perceptual registerings of actual performances with an imagined (perhaps ideal) performance sustained in the imagination. We can also sustain a claim to having a centralised control system *of a sort* in so far as there is knowledge of the state of the system as a whole as well as of discrepancies between the performance of the organism and the imagined template, and this knowledge can interact dynamically with, and thus modulate, relevant systems of motor control by way of the system of cortical projections described by Donald and Deacon. And yet this interaction need not represent an overriding centralised instructive system of control. As with the description of the concert pianist at the conclusion of Chapter 8, we can expect comparisons and corrections issuing from the exercise of kinematic imagination to vary in scale and detail from the overall form of the performance down to isolated finer details. So we can get a form of unified, integrated conscious monitoring and control without the need for a superimposed, instruction-issuing centralised controller, and while still respecting Gallagher's insight that conscious control (and free will) are in general engaged with action at the level of intentions, not of neural firings and muscle activations. The conscious agent engaged in acts of kinematic imagination does not mysteriously reach down into the motor cortex to activate individual motor neurons – instead, the imaginative activities of the agent might be expected, for example, to smoothly interface with ongoing dynamic processes of soft assembling an ever-improving performance that progressively approaches the imagined goal or ideal.

Beyond the Imagined Body in Action

At the outset of this chapter, I set out my intention to explore two less traditional avenues in which we might locate distinctive degrees of freedom for human agents, the first of which involved giving imagination its rightful place in the psychology of our agency. In examining the significance of imagination for agency, besides my brief noting of certain ideas in the work of Dennett, McCrone and McGinn, my focus thus far been mostly limited to kinematic imagination, in part so as to emphasise the case for pre-linguistic roots of our capacity for freedom and, relatedly, for normativity. I think, however, that there is more to be said, and more to be explored, in giving imagination its rightful place within an account of free agency.

Earlier, I cited Colin McGinn's suggestive comments about the centrality of imagination to our free agency. The context of McGinn's comments was his discussion of the stimulus freedom of language – Chomsky's idea[228] that linguistic use, unlike the conditioned reflexes of behaviourist learning theory, does not require the presence of some "ambient stimulus" (McGinn, 2004, p153). McGinn's intention, in terms of the central theses of his book, is to draw a parallel between a language-related ability to speak of the past and the future, as well as of distant, absent and non-existent things, and the imagination's special capacity to deal in "absence, non-existence, revision, [and] outright invention" (ibid).

I see no need to specifically endorse McGinn's thesis that we should try to account for the stimulus freedom of language by grounding linguistic understanding in imagination. I am inclined to think that imagination is critical to language; but at the same time, the relationship might be more dynamic and reciprocal than McGinn allows. Imagination might, for example, 'specialise' in absence and revision, but it may well need the machinery of symbolic representation and thought before it can truly offer to master the never-before-encountered, the non-existent, the impossible, and negation[229].

Whatever the details of the relationship and interplay between the imagination and language, they comprise a powerful combination. Kinematic imagination can present the body in a recalled or an ideal performance; the symbolically-enabled imagination can present as a possibility for action, or as a possible consequence of action, or as a possible future life, or a possible alternative existence, etc. literally anything that can conceivably be imaged, symbolically presented, or symbolically communicated. It is the true playground of the free agent in which degrees of freedom, in terms of possibilities for action, come into their own[230].

In any individual, the capacity of the symbolically-enabled imagination is constrained by a number of variables. Not all imaginations are born equal, nor are they nurtured and developed

[228] See Chomsky (1959).
[229] I am thinking here of CS Peirce's account of iconic, indexical and symbolic reference, and the importance of the distributed and mutually supportive nature of symbolic reference in enabling reference to states of affairs that do not obtain, have never obtained, and even could never obtain. See Deacon (1997) for a useful interpretation and exploitation of Peirce's ideas.
[230] Not that these degrees of (imaginative) freedom can allow the agent to transcend what the Existentialists would have called 'facticity.' No amount of imagining unassisted flight will make it a possibility for action for a human agent.

equally. Some agents experience special constraints – for example, restricted capacity to imagine and foresee the responses and states of mind of others[231], and/or a tendency or restriction to more concrete associations and flights of imaginative projection/ exploration[232]. Other agents might experience an almost unconstrained – perhaps experienced as *uncontained* – capacity for imagination[233]; and some might retreat into the world of their imagination to escape the trauma and the unmanageable challenges that life can hurl in their direction[234].

At the same time that we acknowledge the probability of pre-existing individual differences in imaginative capacity and constraints on the imagination, we should also highlight and encourage the exploration of avenues for growing the imagination. I have in mind, in particular, the much neglected topic of fiction and its potential role in influencing, shaping and adding to the degrees of freedom associated with agency. While fiction technically falls under the category of one of the symbolic technologies discussed later in this chapter, it has a special relationship to the imagination in terms of the journeys on which it (or, at least, the best of it) can lead the imagining agent, expanding their horizons of experience and understanding, fostering their empathic appreciation for the complex lives, experiences and worldviews of others, teaching any number of vicarious lessons, and yet all within a context of suspended belief. The capacity to listen to and read stories, to appreciate and imaginatively immerse oneself in a world of fiction, is one of the truly spectacular achievements of human mentality and imagination, and its implications for our freedom deserve far greater attention and exploration.

More generally, and perhaps less speculatively, there is a wealth of empirical and empirically-inspired work on imagination outside of philosophy that warrants exploration and engagement in pursuit of a more nuanced, sophisticated and empirically plausible account of what our free agency amounts to. I have in mind, in particular, the work of the developmental psychologist Paul Harris, whose extensive research into the imagination of children calls out

[231] I have in mind here individuals on the autism spectrum. See, for example, Baron-Cohen (1995).
[232] Again, individuals on the autism spectrum present possible cases to consider here, especially when it comes to imaginative play. See, for example, Jarrold (2003) for a review of research on the issue of autism and pretend play. But see Leevers and Harris (1998) for important cautions in drawing conclusions about imaginative abilities in autism.
[233] Individuals in florid psychotic states, for example.
[234] Individuals in fugue states, perhaps.

for more detailed and critical consideration within debates about the nature, origins and development of agency[235].

For instance, in the introduction of his book *The Work of the Imagination*, Harris (2000) sets himself the task of arguing for the centrality of *imagining alternative possibilities* within the emerging cognitive capacities of children:

> … the capacity to imagine alternative possibilities and to work out their implications emerges early in the course of children's development and lasts a lifetime. This capacity is especially obvious in children's games of pretend play, but it invades and transforms their developing conception of reality itself. (Harris, 2000, ppxi-xii)

A key feature of the account Harris (2000) develops is the claim that the capacity to imagine alternative possibilities and, specifically, to use this capacity to construct a contrast between what was done and what was not done, turns out to be a critical feature of how children analyse sequences of events, especially sequences and scenarios of human action. That is, according to Harris (2000), children understand actual sequences of events, including the actions of others, in part by imagining contrasting counterfactual scenarios that did not take place. Thus, for Harris (2000), children's capacity for constructing scenes, scenarios and models in their imagination shapes their reasoning and understanding of the world, including notions of causation[236], the actions and intentions of others[237], and concepts of obligation and (even) free agency[238].

---

[235] Somewhat frustratingly, McGinn (2004) acknowledges Harris' work, but he does so in an endnote in which he recommends Harris (2000) as an empirically-informed discussion of the imagination of the child. Given some of their overlapping interests and concerns, it would have been more interesting to see how McGinn might accommodate the experimental and other empirical findings that Harris (2000) presents.

[236] For example, Harris, German and Mills (1996) demonstrated that children as young as 3 years of age were adept at reasoning using counterfactuals, and that they naturally deployed counterfactual scenarios in their causal thinking. Harris' (2000) hypothesis is that these counterfactual scenarios are constructed in the imagination of the child, and form the basis for resulting reasoning and inference.

[237] Harris (2000) mentions the work of Meltzoff (1995) as bearing on the issue of young children appreciating the intentions behind the actions of others. In Meltzoff's (1995) study, 18-month-olds in the experimental condition showed that they could infer the intended act of an observed adult while having only been exposed to the adult's failed attempts to perform the given act. That is, children who had only seen an adult model an intention to act, but fail to successfully perform the target act with a set of objects, were as likely to perform the (unseen) target act as were a group who had seen the target act successfully modelled; both groups produced significantly more target acts than controls who had not witnessed the modelling of either a target act or an intention to perform a target act. Within Harris' (2000) framework, this suggests that that intention group were able to both infer the intention of the adult *and* imagine the successful performance of the target act which they had not seen.

[238] According to Harris (2000, p159), children's concept of a free agent involves someone "who can either carry out or withhold a given action." Harris' developmental evidence would thus suggest that children see the capacity to have done differently as crucial to qualifying as a free agent. To what extent this remains the same over the course of development might be disputed by some – see Nahmias, Morris, Nadelhoffer and Turner (2004).

Of particular interest, given our preceding discussions, is Harris' (2000) view on imagination and emotion, which he develops in part by considering our (puzzling) tendency to respond emotionally to fiction (and fantasy), even though we are aware that what we are responding to is not real and has not happened. Specifically, he asks the question:

> Why is it that human beings, children as well as adults, are prone to display a fully fledged emotional reaction, not just when confronted by an actual situation, but when they imagine a possible situation? (Harris, 2000, p88)

Or, putting it slightly differently:

> Why are we so designed as human beings that imagined inputs readily drive our emotional system? Why is information about whether an event is real or imaginary not automatically deployed to fine-tune our emotional reactions?... in default mode, we are a species that thrills to fictional dangers or sheds tears for imaginary heartbreaks. We may ask… what the biological pay-off might be[?] (Harris, 2000, p84)

Harris (2000) offers two speculative answers to this question, the first of which draws on Damasio's account of somatic markers that we encountered in Chapter 9. The second involves the idea that a crucial function of language, over the course of its (and our) evolution, has been to convey information about the what is *not* here and *not* now – what Harris (2000) refers to as displaced communication or *displaced testimony*.

In terms of his first speculative answer, Harris (2000) in effect reinterprets Damasio's (1994) account of somatic markers through the lens of employing imagination during deliberation and decision making: decision-making is an embodied, 'warm-blooded' affair involving real generation of emotions because, when we imagine what we might do, we are able to experience emotions of a kind with what we would feel if we go ahead and do it. This would suggest that perhaps Damasio is incorrect to view Elliot and other VMPFC patients as merely suffering from a disruption to their secondary emotion system: if Harris is correct, perhaps the deficiency is one that also involves a failure of imagination.

Harris' (2000) second speculation touches on the value of testimony and the communicative power of symbolic language. In its ability to communicate information about the absent and the past – the domain of testimony by another who was there – language offers the opportunity to learn at one or more removes from first-hand experience. In this context of evolving language and communicative practice, Harris (2000) asks us to consider two things: first, it is likely that the kinds of events that will be relayed via honest testimony will often have included emotionally charged events; and, second, that a failure to generate real and appropriate emotion when an emotionally charged event was being relayed would leave the

audience both lacking the knowledge of the emotional dimensions of the event (emotional learning would only take place in first hand experiences) as well as socially (empathically and pragmatically) out of sync with the person whose testimony is being heard. Whereas:

> …if our interlocutor's emotionally charged eye-witness report kindles in us the same emotion as the events themselves aroused in him or her, we end up being in tune with our interlocutor and alerted to events beyond our immediate horizon. (Harris, 2000, p90)

Harris' (2000) proposals are speculative and, at least in the case of the somatic marker hypothesis, probably open to empirical testing. But they once again point to potential avenues for exploring the full significance of imagination to human agency, both as embodied, emotional beings who must deliberate and decide in a suitably warm-blooded fashion, and as social creatures whose possibilities for agency must be understood within a context of what Donald calls a cognitive community and that, following Harris' lead, we might want to subtly reframe as a community of shared ideas and experiences, both real and imagined.

There is much more to say in developing these ideas about the significance of imagination to human agency and human freedom. Giving the imagination its full and proper recognition as a critical element in our embodied mental economy promises us new directions for exploring and securing claims of origination, novelty, ownership and control that are the basis for various degrees of freedom we wish to claim as free agents. At the same time, because imagination (as discussed here) involves the conjuring and constructing of images in extended consciousness, exploring and articulating the connections between imagination and agency promises to leave behind the spectre of the Agent Automaton as a sceptical (and, one is tempted to add, behaviourist) chimera; while the grounding of some of our highest cognitive and motor achievements in *imaginative* activity, rather than some kind of mental (or neural) logical calculus, promises some respite from the dangers of presuming ourselves to be Hyper-rational, hyper-reflective agents.

The ideas presented here are, intentionally, suggestive and programmatic – they outline future paths to explore outside the walls of the traditional debate, paths that will be welcomed by those who, like me, have come to the conclusion that there must be more to be said about free agency than seems to be allowed within the topography of an impasse. And, as outlined in Chapter 5, I think that even these programmatic suggestions about the significance of imagination can be translated into an hypothesis about where an incompatibilist account of

free agency (that is not a traditional libertarian account) might try to locate a form of indeterminism that is friendly to, if not required for, free agency. But before concluding with a discussion of that hypothesis, I first want to explore the second of the less-explored avenues I gestured towards at the beginning of the chapter – ideas about various externalised aspects of mind, and the significance might have in human agents being able to transcend various biological constraints on their agency.

Individual Agents: Limited Capacities and Resources

In Chapter 9, we encountered Damasio's hypothesis that emotion is critical to ordinary processes of deliberation. Part of the evidence and logic behind this hypothesis came from the speculation that patients like Elliot suffered from a subtle working-memory related deficit when they engaged in the real world (as opposed to laboratory) dynamics of open-ended deliberation. As Damasio (1994) makes clear in his discussion of the extensive testing that Elliot underwent, the deficit was not one of working memory *per se*. Instead, according to Damasio's hypothesis, the demands of open-ended deliberation in the absence of emotional/ somatic marking very quickly run up against and then exceed the capacity of even a normal working memory.

It is obvious that individual human agents have all sorts of resource and capacity limitations that act as constraints on their possible avenues of psychological and bodily activity. Sceptics about consciousness are, for example, especially keen to enumerate the limitations of conscious task performance under a variety of laboratory conditions[239]. A relatively simple example that illustrates the point is offered by Andy Clark (1997). Given a certain amount of learning in school, most of us are capable of instantly producing an answer to simple multiplication problems, such a '9 x 7 = 63'. However, the majority of us would find that '916 x 753' exceeds our powers of mental arithmetic. Instead, we would have learned a method of externalising the cognitive task at hand, writing down the problem in a conventional notation, and then tackling it in a sequential, rule-governed manner, such that the problem is broken down into a series of simple multiplication operations, a procedure for factoring in multiplication by tens, hundreds, etc., and finally the summation of our smaller component answers to reach the final answer.

[239] See Donald (2001, pp14-25) for a useful summary of some of these capacity limitations.

This example, though clearly an instance of a formulaic externalised cognitive operation, points to a number of interesting features shared by a range of human cognitive activities. First, the context of externalising the cognitive task at hand is one in which the processing capacity limitations of the human mind, for all its remarkable power, are very quickly reached. Second, perhaps the most significant kind of capacity limitation that we run into under such circumstances is that of the biological memory systems. Under normal circumstances, the underlying reason why we cannot mentally calculate '916 x 753' is that we cannot realistically hope to memorise our times tables all the way up to our 916 times table (that is, in addition to the fact that this would be a monumental waste of time). Nor are our memory systems well suited to breaking down the calculation into, for example, three segments of '900 x 700', '16 x 50', and '16 x 3', because interference effects are likely to occur in trying to hold the answers to each segment in short-term memory while performing the other calculations. (Such a method may, needless to say, also run into problems of processing limitations, depending on an individual's mental maths ability, as well as the complexity of the calculation.) Long multiplication provides a reliable, externally-scaffolded method by which someone who has mastered the technique, who has memorised the times tables up to 9, and who is capable of long summation (another externalised cognitive technique), can tackle at will multiplication problems of almost any length and complexity.

Third, according to Clark (1997), it would appear that to the extent that there are people who can carry out operations like '916 x 753' mentally, it is by way of them having internalised the procedures of external long multiplication. They are able to utilise their capacities for mental imagery, along with working and short term memory, to follow the same steps as the externally-scaffolded procedure, but without recording the steps or their outcomes in an external medium.

Fourth, the central reason why there are few who do, and quite probably few who can successfully internalise and utilise the long multiplication procedure for operations as complex as '916 x 753', is that the working and short term memory demands involved in holding all the relevant component answers 'in mind' while performing other parts of the calculation, and the likely interference effects that can arise, exceed those of the average human subject – certainly those who have not received some form of special memory training. But, fifth and finally, it seems rather pointless to go to all the effort of internalising the technique, given the speed, efficiency and reliability of the externally scaffolded

procedure. We can, in deciding (as learners or as educators) how best to utilise our limited processing and memory resources, reflectively apply the '007 Principle'[240] to ourselves, and recognise that developing an internal version of the long multiplication method is for the most part a monumental waste of time when simple interactions with, and manipulations of the external environment will yield more reliable (and probably more speedy) results with considerably less cognitive effort. (And, unlike the internal version, the results and procedure are thereafter available for checking by ourselves and others.)

Upon reflection, we can find externalised cognitive procedures in just about every aspect of the life of literate persons the world over. We make lists of groceries to buy and things to do as aids to our memory. We sometimes use lists or other external records of wishes, plans and intentions to examine these in a more holistic way than the serial nature of conscious reflective and deliberative processes ordinarily allow, comparing, prioritising and editing these 'to-do-lists' as we go. We entrust memories of phone numbers, appointments, names, and other items to various external media such as note pads, diaries, and (increasingly) digital media such as cellular phones, PDA's, digital voice recorders, and computers. Some use external media to record personal experiences, thoughts, feelings and reflections in the form of a journal or diary; and some do the same with thoughts, feelings and dreams as a specific aid to reflection on their own psychological make up. And the working world, outside of manufacturing and physical labour, is literally filled with a range of external records and memory devices, as well as externalised and distributed (in the sense of inter-individually distributed) cognitive processes.

At the core of most of these cognitive processes lie a variety of symbolic technologies – primarily writing and similar symbolic notations – whose significance for understanding modern human cognition cannot be underestimated:

> Symbolic technology is the enterprise of manufacturing and crafting external symbolic artefacts and devices. These have enabled us to build a vast cultural storehouse and an external symbolic storage system, which serves as a permanent group memory and includes such things as books, museums, measuring instruments, calendars, and computers. These are extensions of what archaeologists call material culture. But unlike most aspects of material culture, they are designed specifically to help us think, remember, and represent reality… They revolutionise what we can do with our minds. (Donald, 2001, p305)

[240] "In general, evolved creatures will neither store nor process information in costly ways when they can use the structure of the environment and their operations upon it as a convenient stand-in for the information-processing operations concerned. That is, know only so much as you need to know to get the job done." (Clark, 1989, p64)

Donald (2001) thus sees the impact of symbolic technologies as lying both at an individual and at a cultural level. These technologies are a cultural phenomenon, and in so far as they are symbolic, are dependent on a variety of historical and ongoing social features of the contexts in which they operate. And the collective storage and sharing of knowledge and ideas enabled by these symbolic technologies have dramatic consequences for the nature and possibilities of human societies.

But the impact of these technologies is equally important at an individual level – at the level of what individual minds are able to do as a result of the presence and use of these external media. Symbolic technologies allow us to escape the biological limits of our brain's various memory systems, however we choose to classify these[241]. But the significance of this externalisation of memory does not lie solely at the level of increased (even limitless) capacity. Externalising memory – creating what Donald likes to call the 'external memory field – also changes the way in which the contents of memory are available to consciousness, and with such changes come the possibilities for new cognitive strategies:

> [The external memory field] is an extraordinary historical development because it changes the long-standing relationship of consciousness to its representations. We can arrange ideas in the external memory field, where they can be examined and subjected to classification, comparison, and experimentation, just as physical objects can in a laboratory. In this way, externally displayed thoughts can be assembled into complex arguments much more easily than they can in biological memory. Images displayed in this field are vivid and enduring, unlike the fleeting ghosts of imagination. This enables us to see them clearly, play with them, and craft them into finished products, to a level of refinement that is impossible for an unaided brain. Thus the display characteristics of the external memory field expand the range of mental operations available to a conscious mind. (Donald, 2001, p309)

Donald is, in essence, claiming that the list of externalised cognitive activities that I mentioned above in fact reflects activities that considerably expand the power of the individual mind in ways we may simply take for granted, having grown up in a literate culture saturated with, and heavily dependent on, such external storage devices.

Donald's (2001) claims regarding the external memory field suggest that the precision, stability and controlled malleability of the external memory field allow us to carry out whole varieties of cognitive experiments, the capacity demands of which far outstrip the resources of an isolated mind-brain restricted to working within the workspace of conscious working memory.

[241] For purposes of this discussion, a standard division into long-term, short-term and working memory systems will serve adequately.

The power of the external memory derives in large part from the *form* of external memory itself:

> [E]xternal symbols [exograms] give us stable, permanent, virtually unlimited memory records that are infinitely reformattable and more easily displayed to awareness. Moreover, exograms are much easier to search [than internal biological memories], and we can recall them with a variety of retrieval methods. The availability of powerful external media [therefore] increases the number of ways we can represent reality. (Donald, 2001, pp309-10)

While we should in no way think that these claims imply a dismissal or devaluing of the cognitive and behavioural significance of imagination, conscious reflection, planning and deliberation, we should recognise that Donald is trying to point us to a different *order* of imagining, conceiving, planning and understanding that is only made possible through the addition of external memory media to our psychological toolbox. If, as Donald claims, we can dramatically increase the numbers of ways in which we can conceive of and represent reality, then we can also infer a dramatic increase in the ways in which we can conceive and plan our actions in the world.

With our increasing dependence on digital media, the astonishing growth and size of things like the Internet, and the incredible speed of digital search procedures (think of how long the average Google search takes), the possibilities for effective storage, search and retrieval are constantly evolving into ever more powerful cognitive tools[242] (even if the parallel explosion in the sheer quantity of information available to an individual undermines some of these advances to a significant degree, and creates real problems of 'information overload', not to mention inundation with useless or low quality information).

The availability of, and interaction with, the external memory field transforms the internal processes of mind. Donald (2001) describes what he thinks of as a mirroring arrangement between the internal and external memory systems, with the architecture of biological memory being externally captured in the symbolic environment, and this image then being reflected back into the mind-brain to interact in awareness with the biological memory systems themselves. This interaction, he claims, has notable implications for the mind:

> This mirror arrangement… changes the reflective power of the conscious mind, because the external memory field gives working memory a much more solid display system for representations. Ordinarily, neither spoken words nor images can be displayed long enough in

[242] An example close to home – the crude surveys on the occurrence of the term 'imagination' at the beginning of this chapter would have been practically impossible, and certainly prohibitively time-consuming, in the absence of electronic access and search functions (in this case, via Amazon).

> normal working memory, even in the intermediate term, to allow any prolonged or detailed reflection on them. Natural memory display is poor, lacks definition and detail, and is notoriously unreliable. The external memory field gives us sharper and more durable mental representations. This allows the conscious mind to reflect on thought itself and to evolve longer, more abstract procedures that serve to verify and control the quality of its own actions. (Donald, 2001, p313)

This arrangement leads, in turn, to greater capacity for abstraction, for turning thinking into a more formalised and systematically developed activity, the products of which can once again be externalised, stored and reflected back into the mind-brains of individuals to thereby transform the very processes of cognition. Even in brief outline, this account highlights the dramatic potential of externalised memory systems (and other symbolic technologies) to impact on the (superplastic) mind-brain systems on which these technologies depend for their existence and significance.

Clark (1997), who is sympathetic to Donald's theory (at least as put forward in Donald, 1991), emphasises a different aspect of the interaction between the mind-brain and the symbolic memory traces in the external memory field[243]. Clark is partial to neural network approaches to modelling the mind, in part because he thinks that the pattern detection and completion capabilities displayed by neural network simulations model an especially important and powerful feature of biological mind-brain systems. That is, Clark locates much of the processing and resolving power of embodied mind-brains in their ability to detect, generate, respond to, and complete patterns in sensory input, generating associations and gestalt perceptions in the process. It is not hard to see how such an emphasis on pattern detection, generation and completion can point us towards a significant form of interaction between the external memory field and the perceptual-cognitive apparatus.

To begin, we can illustrate Clark's idea by considering an anecdotal example. My sister, a teacher by profession, is fond of using word-formulation and word-detection games in class to improve vocabulary and spelling. Word-formulation games follow rules like those of the board game 'Scrabble', with a learner having to formulate words from randomly selected letters of the alphabet; word-detection tasks may involve clues, conundrums or riddles, with the learner having to rearrange a given set of letters into the correct word. My sister found in her own case, and then quite generally in the case of her students attempting such tasks, that arranging or writing the letters in a circle both speeded up the rate at which such tasks could

[243] Clark's work on the externalized and extended mind has continued up until the present. Notable works on this topic include his collaboration with David Chalmers on something of a 'policy statement' of a paper titled "The Extended Mind" (Clark & Chalmers, 1998), and his recent book (Clark, 2008) titled *Supersizing the Mind*.

be performed (especially the word-detection tasks), and also enabled greater word productivity in the word-production tasks. While I have not scoured the cognitive psychology literature to find more rigorous evidence of such an effect, the anecdotal evidence is suggestive of the kind of effect that Clark has in mind – namely that, by externalising and physically rearranging symbol tokens, and then focussing the resolving and pattern-detecting powers of the visual processing system on these tokens, solutions to what is otherwise clearly a literate, cognitive task can be more readily and speedily produced through the interaction between the perceptual and language systems in the mind-brain. And we might expect similar effects to emerge in the domains of thought, deliberation and action more generally.

Consider, for example, the contrast between a conscious subject mentally examining or considering, in relatively serial fashion, a list of intended actions to be completed in the day ahead. In so far as this thought process takes place by way of inner speech, then what Donald has referred to as the literate parts of the brain may plausibly be imagined as being active in such a process, and similarities, conflicts and other patterns in the list of actions may or may not be detected, depending on the systematic nature of the thought process, and the adequate functioning of working and short-term memory systems. If we now consider the list having been externalised by writing items down on a notepad, we can plausibly imagine that the ability to detect similarities, conflicts and other patterns will be enhanced because the sophisticated resolving and pattern-detecting characteristics of the perceptual apparatus can be put into service alongside deliberations (in inner speech) going on in awareness. Thus, not only is memory enhanced through the availability of the external memory field (and, with it, possibilities for more formal, explicit search and comparison procedures), but the externalisation of thought contents can itself bring the processing powers of the perceptual-cognitive apparatus to bear on those contents in ways that differ significantly from what is involved in internal thought and deliberation. Furthermore, the suggestive example of word-production and word-detection tasks points to the possibility that, while Donald is quite justified in emphasising the degree to which the externalisation of thought enables more formal, explicit, rigorous and deliberate procedures, Clark's emphasis on perceptual processing allows for a less deliberate, and in this sense less conscious, set of procedures whereby externalisation enhances cognition. That is, we need not always know why externalisation aids us in finding cognitive solutions, or what the precise mechanisms and features are of the perceptual-cognitive systems that aid us in producing these solutions. We may simply have found, individually and/or collectively, that externalisation 'works for us' if

we follow certain procedures, and such discoveries are sufficient to entrench a range of interactions with symbolic technologies and the external memory field that broaden our cognitive and behavioural horizons.

Symbolic technologies, the externalised memory field, and externalising and extending of the mind's capacities for thought, imagination, planning, reasoning, and any number of other mental exploits – these help constitute the second avenue for exploring and claiming truly novel degrees of freedom in human agency. As with my proposals relating to imagination earlier in this chapter, this is an area that has not featured significantly in traditional discussions of free will; moreover, it is an avenue of investigation that promises to deliver degrees of freedom in agency while avoiding both the spectre of the Agent Automaton (Donald's account, in particular, being one that emphasises the consequences of externalising the mind for the structure and capacity of *consciousness*) and the risks of pretending we are, or aspiring to be, hyper-rational, hyper-reflective agents (much of the value of externalised symbolic technologies being so easy for us to recognise precisely because of our limits as thinkers and agents). It remains for future projects to deliver on this promise.

Imagination, Agency and Indeterminism: A Speculative Hypothesis

Before closing, I must return to an issue raised in Chapter 5, where I offered the beginnings of a response to the question "Where, given my incompatibilism, do I see indeterminism fitting into a picture of free agency?" While my initial two responses sought to deflect the question as either a matter on which I need not say anything in particular, or (more likely) a matter that I could only address to any degree of satisfaction after taking on a much larger project of displacing reductionistic physicalism along with its various truisms and habits of thought, I did outline one positive, speculative hypothesis concerning indeterminism and imagination, with the promise that I would return to this hypothesis at the end of the thesis. Having reached that point, I must now deliver on that promise.

As will be recalled from Chapter 5, the key challenge in proposing and developing such a speculative hypothesis about a possible place for indeterminism *within* the psychology of a free agent is that such indeterminism should not threaten the integrity, ownership or control of the agent in the ways I have (in Chapter 2) alleged for traditional libertarianism. If I am right in claiming that the libertarian tactic of inserting indeterminism into moments of choice and volition is a failure, is there any viable alternative?

My speculative and programmatic answer is that indeterminacy within certain processes involving *imagination* could be recognised as an important ground for our claim to being free agents without thereby threatening the integrity of the agent. As we have seen in the current chapter, both McGinn (2004) and Harris (2000) give imagination pride of place in providing the space in which we generate, contemplate and test out the consequences of alternative possibilities for action. We deploy our imaginative capacities not only in imagining alternative possibilities for our action, but in analysing the actions and intentions of others (even when, as in Meltzoff's (1995) study, they fail in their execution of their intended actions). And we employ kinematic imagination (Donald, 2001), and imagined alternative possibilities for the form of an action, in acquiring and shaping skilled performances. To the extent, then, that our degrees of freedom as human agents depend on our generation of alternative possibilities for action, imagination is both playground and testing ground for the free agent.

At the same time, the realm of the imagination is rich with associations of creativity, novelty, origination and ownership – concepts that have important associations with free agency over and above notions of alternative possibilities. And it is in the context of the generativity of imagination – the possibilities for genuine novelty and origination, for imaginative products that truly transcend experiential (and other) inputs – that I propose we explore finding a place for indeterministic processes that boost *free* agency whilst not at the same time undermining the claims of ownership and control over choice and action that the demands of agency would have us satisfy first *before* qualifying as free agents.

How might this work? My proposal involves adapting two ideas from Kane.
First, Kane proposed to help ground claims of Ultimate Responsibility by positing occasional, special moments of self-shaping – his SFAs – in which undetermined events within the agent might come to shape their future reasons and decisions, determined or not. Indeterminism need not be inserted into each and all instances of choice in order for these to be the choices of a free agent. It might be sufficient that, in free agents, we could in part trace their choices back to these special, undetermined moments of self-shaping. Second, particularly in the context of defending Kane against Dennett's (2003) criticisms, I allowed that it could matter,

in principle, both that an indeterministic process of self-shaping occurred *inside* the agent[244], and that it could be significant that the indeterminacy involved was attributable to the psychology of the agent[245].

Suppose, then, that there are important moments of self-shaping that occur in agents, both in early development and across the lifespan. *Contra* Kane, these moments of self-shaping are not Self-forming Actions (SFAs) involving choice between competing alternatives under conditions of plural voluntary control. Instead, like the impact of a poem or a work of fiction on an individual consciousness, or the creative maelstrom that gives birth to a work of art, these moments of self-shaping activity are moments of *imaginative generativity* in which options and possibilities for the future are conjured and constructed as scenarios in the imagination. By hypothesis, these moments of imaginative generativity are indeterministic in some significant sense. The possibilities and scenarios imagined emerge as a complex probabilistic function of (amongst other potential variables) the psychology, psychosocial history, and brain dynamics of the individual agent.

Contra Dennett, it matters a great deal that the indeterminism involved in these moments of imaginative self-shaping is internal to, and indeed a function of the history and internal dynamics of the agent. These works of the imagination represent the creative, novelty-producing confluence of all the influences in the individual agent's life – their experiences, ideas, emotions, values, hopes and fears; their history of decisions, projects, actions, successes and failures; their second and third hand experiences imagined on the experience and testimony of others; their experience and know-how and knowledge gained in the make-believe worlds of pretend play, fantasy, fiction, poetry and art; and their parental and societal context (to name just some of the potential variables). The probabilistic and chaotic dynamics that I speculate will be at play in these moments of creatively imagining options for the future will be as uniquely individual as the pattern of neural connections in the brains of even genetically identical individuals[246]. For this reason alone, the internality and ownership of

[244] As compared to Dennett's (2003) idea that being connected, by remote, to a geiger counter would achieve the same ends in resolving the tension in a SFA.
[245] As compared to the anti-libertarian tradition, well-represented by Dennett (2003), of insisting that randomness is just randomness.
[246] See Chapter 8, and (for example) Edelman (2004).

these indeterministic dynamics matters a great deal as an expression of who that individual is, and it matters also to the possible creative outputs of the process[247].

There are further important differences to note between this proposed idea of imaginative self-formation and the account of SFAs offered by Kane. Unlike SFAs, occasions of imaginative self-formation *are not choice situations*. What emerges from these indeterministic moments in an agent's life are *options for possible futures*, amongst which the agent can make choices. In this way, my proposal attempts to avoid the libertarian move (and resulting anti-libertarian critique) of inserting indeterminism into moments of volition. My proposal involves no particular commitment, at least not at this programmatic stage, as to how choices might be made between imagined possible futures.

Furthermore, because mine is not an account of self-forming choices, it is also not a proposal that would look to make or ground claims of Ultimate Responsibility (*a la* Kane) in these undetermined moments of imaginative generativity. From an incompatibilist perspective, it is enough that such moments can make a meaningful contribution to the possibilities for choice and action in an agent while at the same time breaking any backwards-extending chain/s of deterministic causation that might be claimed to run through an agent's lifeline[248].

Finally, because the proposed moments of undetermined self-shaping are not choices, it is sufficient that the agent can claim ownership of the process and its products. Questions of *control* are not central to either the significance of these moments of self-shaping, nor to the claims of ownership that can be made. When a work of art flows from the imagination of an artist, we do not query their ownership of the product on the basis of quibbles about the extent of control they might have in the in midst of the creative process.) For this reason, any ceding of control that might be thought to come with the activity of indeterministic dynamics does not present the kind of problem for agency that it does in the case of Kane's SFAs (or, for that matter, in an agent-causal account of the moment of choice).

[247] To be fair to Dennett (2003), his argument against Kane was focussed on the idea of indeterminism playing a role in a knife-edge choice situation between a restricted range of known alternatives. I would hope that, in the context of a multidimensional and dynamic process of imaginative generativity, Dennett would not propose that remote radioactive decay could do the same job as internal brain dynamics.
[248] To borrow Rose's (1997) evocative term for the life cycle of an individual organism.

In summary, then, the core of this speculative proposal derives from the idea that in the case of certain imaginative processes, and unlike processes such as reasoning, deliberation and choice, we can *welcome* the open-ended generativity and possibilities for novelty that could result from a degree of indeterminism *without* thinking that this probabilistic (or, if you prefer, chance) element within the process somehow undermined claims of origination, novelty, creativity and ownership on the part of the agent. Open-ended, indeterministic moments of imaginative generativity can be seen as a virtue and an aid to free agency, not as some kind of shortcoming. Conversely, we might make a case for the freedom of an agent being restricted, curtailed or otherwise undermined when there has been a foreclosing on the generativity of various imaginative processes at play in processes of agent shaping. Such foreclosing could be either internally or externally driven; and the tendency to foreclose on the imagining of possible futures that needs to be fostered and protected by the agent and those around them (including their parents and significant others).

(Inevitable) Objections

While I would prefer to leave this speculative hypothesis as a project for future research, it is inevitable that a range of objections will be levelled at it, most notably by compatibilists. I will restrict myself to considering just two of the most inevitable objections. First, in what way does my proposal escape the second horn of the sceptical dilemma posed by Lipton (2004) in Chapter 1 – specifically, how is my proposal for where to fit in indeterminism supposed to remove the worry that, if things are undetermined, then they are not up to us. Second, how will my proposal fair in the face of compatibilist worries over libertarian luck; or, to put it differently, what can I say to defend my proposal against the complaint that different outcomes under conditions of repeated re-runs of these episodes (or in different possible worlds) must surely threaten the value – the freedom-relevant value – of both the process and the outcomes.

My responses will be brief, both because the proposal is still programmatic, and because the responses I have to offer to each objection are very similar in substance. The question of whether or not events that are undetermined can be up to us is, in the context of my proposal, ambiguous. In the sense of ‘up to us-ness’ that requires an agent to have a requisite amount of control over how things turn out, I have already allowed that we do not have such control over the process and outcomes I have described. But this is as it should be – when it comes to open-ended imaginative generativity, what we want is for imagination to ‘run free’. On a

second sense of ‘up to us-ness’, the process and outcomes are up to us because, in so far as they represent the dynamic confluence of a unique set of variables in the agent’s lifeline, there is a very strong sense in which the outcomes are an expression of who that agent is. Moreover, as suggested earlier when comparing moments of imaginative generativity to the production of a work of art, the products of these moments of constructing imagined futures need be no more or less up to us than the artist’s expression of themselves in their art.

With regard to the second objection, we should immediately note that much of the sting in this objection has been taken out by the fact that we are no longer imagining different outcomes to a choice situation where the agent has the same reasons for acting in each case. Instead, we must imagine an agent imagining different possible futures, given the same confluence of variables in their lifeline. Would this make our agent a ‘victim’ of luck? It is not immediately obvious that we should view the case this way. Consider, again, the artist and their creative act of producing a work of art. Supposing this creative process is similarly indeterministic, would we regard the actual piece that emerged as being a (problematic) product of luck because, given the same mix of creative and personal inputs, it could have turned out differently? It is not obvious that we would see a problem here, nor that we would view the art as a mere product of luck. As in the case of the previous objection, I think this response ought to be persuasive.

But perhaps I am asking too much of the analogy with creative, artistic processes. Suppose we consider two instances of imaginative generativity in an agent and their counterpart in suitably close possible worlds. In our world, the end result of the process is the agent having imagined possible futures *a*, *b*, and *c*; while the counterpart only imagines *a* and *b*. Is it merely a matter of luck – and is it somehow a problematic matter of luck – that our agent in this world imagined option *c*?

One response, at this point, would be to remember Kane’s proposal for some ‘conceptual therapy’ regarding the notion of causation: it is a mistake to associate ‘indeterministic’ and ‘undetermined’ with ‘uncaused’; in the case of indeterministic causation, we do not have an absence of causation, but the presence of nondeterministic causation – it is causation nevertheless. Our agent imagining *a*, *b*, and *c* was caused indeterministically by the confluence of unique variables in their lifeline; in the case of their counterpart, their imagining futures *a* and *b* was caused indeterministically by their (identical) confluence of

unique variables in their lifeline. Is this a difference of luck? No, it is a difference in outcomes of an indeterministic causal process where the relevant probabilities were such that either outcome was possible, given the causal influences at work.

If the critic is still not satisfied, there remains the possibility of biting the bullet and accepting the difference between our agent and their counterpart as one of life's little accidents – like the accident I mentioned right at the outset of this project, when I stumbled upon the articles by Bargh and Wegner and their colleagues in a batch of journals discarded by a departing professor. In a less than hyper-rational world, it is alright to accept and embrace the existence of life's little accidents. At least, I imagine that it is ok, as long as we are dealing with options for the future, and it is still up to us what we will make of them.

Chapter 11

Conclusion

This research has sought to establish a number of important conclusions about free agency, our prospects for understanding what it is, and for defending the claim that we, as human agents, can make a meaningful claim to it. Each conclusion makes a contribution to the particular issues at stake, while also making a contribution towards the development of the overarching thesis. I will briefly overview the central conclusions established over the course of the last ten chapters, before making some final remarks about the topics and avenues for future research.

While a number of important conclusions were put forward and defended in Part I (Chapters 1 to 4), the overarching conclusion that provided the immediate impetus for Part II was that the traditional free will debate leads to an impasse that is detrimental to the prospects of our attempts to articulate a plausible and defensible account of the nature and limits of free agency. While there is much of value that can be gleaned from the traditional debate, and while the value of so much careful and disciplined philosophical reflection about free will should not be in any way underestimated, the debate suffers because the main opponents who would want to defend our claim to having free will (compatibilists and libertarians) each try to find something that they are not going to find. As argued in the case of Robert Kane (Chapter 2), libertarians are not going to find free will by trying to insert indeterminism into critical moments of volition. In this sense, the compatibilist suspicion about libertarianism turns out to be fundamentally correct. And yet, as I argued in my case against compatibilism (Chapter 4), it also turns out that the libertarian suspicion about compatibilism is correct: despite all appearances and arguments to the contrary, you cannot get free will in a deterministic universe.

The arguments offered against Kane's libertarianism, and against the global prospects for a sustainable compatibilism in a deterministic universe, stand in their own right as contributions to the very traditional debate that shapes and motivates them. But, as argued in Chapter 5, the outcome of identifying a clear impasse, with no obvious prospect for movement and progress, is not sufficient in itself; and yet nor should it be thought to motivate

hard incompatibilism or apathy. Whether or not we are free agents matters a great deal, and the improbability of decisive movement within the confines of the traditional debate does not make any of the issues at stake any less pressing. Nor should the impasse motivate hard compatibilism, because this would require an additional argument: an argument to the effect that we have no meaningful way of framing either what we mean by, or why we might worry about, the claim that we are (but might not have been) free agents. There are both pressing and promising questions about free agency that call for our attention independent even if addressing them will not (immediately) advance the cause of any entrenched traditional position.

As argued in Chapter 5, an alternative framework can be delineated in which both our concerns about free agency, and a project of articulating and defending positive claims about our freedom, can be situated. That is, we can outline a framework in which we to pursue the subject matter of the traditional debate – the description and defence of free agency – while rejecting, or successfully moving outside of, the traditional framework that some might think of as the source of the content and meaning of this subject matter.

We don't need the parameters of the traditional debate in order to pose sensible and pressing questions about the nature and limits of free agency because we can discern an image of a kind of agent – the Agent Automaton (AA) – whose spectre seems to threaten and undermine our claims to free agency independently of (global) metaphysical doctrines about causation, determinism and indeterminism. Moreover, using this image, we can reframe the central challenges involved in articulating and defending an account of free agency as that of successfully *avoiding* the spectre of our being AAs while, at the same time, *avoiding* the trap of trying to secure free agency by asserting (or hoping) that we might be some kind of empirically implausible super-agent – what I call a Hyper-rational, Hyper-reflective agent (HHA). Resituating the debate in this way promises to reinvigorate research into free agency, with the prospect for saving important insights from both sides of the traditional debate, without becoming entangled once again in a debilitating impasse.

Having thus recast the debate about free agency as the challenge of defending against claims that we are AAs while avoiding the trap of insisting that we are, instead, HHAs, I argued (in Chapter 6) that psychology and allied empirical sciences of the mind and brain have furnished us with empirical data and empirically-inspired hypotheses relating to agency that

some have interpreted as grounds for concluding that we are indeed, as it turns out, AAs (or, more cautiously, that we function much more like AAs than we might like to think, to the extent that if we give due recognition to these aspects of our agency, we are unlikely to be able to sustain a case for our being free agents). The bulk of this evidence can be roughly divided into two: on one hand, there is a range of evidence taken to suggest that we have far less conscious control over our mental lives, our choices and our behaviour than could be consistent with any strong claim that we are free agents; and, on the other hand, varieties of evidence have been highlighted that suggest we are regularly (and thus, potentially, are in general) wrong in our own claims about, and interpretations of, our motivations, intentions, and conscious involvement in the initiation and control of behaviour. Taken together, this evidence can be used to mount a challenging sceptical case to the effect that we are, in fact, agent automatons who operate under various illusions as to the nature of our own agency. As I argued in Chapter 6, such a case deserves a response. It deserves a response that is neither dismissive nor hamstrung because of immobility within the traditional debate. Indeed, constructing and highlighting this sceptical challenge is, in itself, an attempt to partly relocate the problem of free agency as a problem in psychology, thereby injecting new perspectives, new challenges and new data for our consideration.

Thus, the sceptical case deserves a response that is equally adept at addressing both the more philosophical as well as the more empirical issues raised by the evidence and arguments on offer. The balance of the thesis has been concerned with taking up this challenge, while at the same time developing and/or pointing out positive directions in which to take the construction of an account of free agency that is no longer defined and constrained by the framework of compatibilism-versus-incompatibilism.

In Chapters 7 and 8, I mounted a defence of free agency against these empirically-inspired sceptical attacks, drawing on a combination of philosophical and empirical/ psychological resources. The central conclusions for which I argued can be summarised as follows:

a. The presumed ontology and timescale in which conscious agency has been examined and analysed in these sceptical arguments falls foul of important philosophical correctives and clarifications about mental states, intentional action, and consciousness. In short, the sceptical empirical arguments present attacks on conscious agency in forms, and on timescales, that are not the proper target for attributing (or understanding) either conscious agency or free agency.

b. There is more than just one variety of control that we should expect to find operating in complex intentional biological systems such as human agents. The sceptical empirical arguments depend too strongly on a simplistic, and empirically implausible, assumption that conscious control of intentional behaviour will fit only one model or image of control.
c. To the extent that we might need to identify, describe and defend more distinctively human forms of (conscious) control in order to sustain claims of our being free agents, the evidence cited in the sceptical arguments is not adequate to the task of undermining claims that we exhibit highly adaptive and efficacious capacities for intermediate-term conscious governance.

At the same time that these conclusions sought to substantively address the concerns raised by the discussions of Chapter 6, the *method* of addressing them also demonstrated the value of relocating the debate to within a problem space shaped by psychology and allied interdisciplinary sciences of the mind.

The partly defensive and partly constructive case for the importance and efficacy of conscious human agency offered in these chapters was followed, in Chapters 9 and 10, by a series of arguments against the risk of reifying and/or over-intellectualising our capacity/ies for self-governance – i.e. turn ourselves into HHAs – in pursuit of an account of a distinctive and free form of human agency (a risk not taken seriously enough by many accounts of agency located within the traditional debate). As in the arguments of Chapters 7 and 8, the arguments and conclusions of Chapters 9 and 10 represent an attempt to integrate philosophical and empirical insights into the nature of human agency within an interdisciplinary problem space, saving and integrating insights from the traditional debates on free will while being freed from many of its constraints. The central conclusions of these last two chapters include:

i. We need to recognise that empirical evidence regarding the probable limits of conscious, reflective deliberative agency, specifically with regard to our capacity for self-insight, as well as the potentially deleterious effects of excessive conscious reflection and articulation (that is, the possible risks of thinking too much). As in (ii) below, our conscious capacity as agents is at its best when engaged with the world, rather than when introspecting and articulating our activities in and responses to it.
ii. A compelling combination of empirical (Damasio) and philosophical (Blackburn) observations and arguments can be used to sound a warning against the error of

detaching reasoning and deliberation from the feeling, acting body of the agent engaged in that reasoning/ deliberation: it turns out that the cool, dispassionate reason separated from emotion is better associated with real-world pathology and lives in disarray rather than with some ideal of rational and autonomous agency. The deliberative capacities of a free human agent in the real world are not introspective, hyper-rational cognitive powers stripped of the push and pull of affective life. Instead, these are remarkable and complex world-focussed capacities for self-governance shaped, coloured and, sometimes, distorted by the animating force of emotional states.

iii. A compelling empirically- (and evolutionarily-) inspired case was made to associate important dimensions of reflexivity and normativity with embodied capacities for kinematic imagination, skill acquisition, and conscious governance of skilled performance. Such associations would act as a corrective or balance against the tendency to draw too heavily on linguistic, cognitive and logical operations, especially those associated with the propositional attitudes and practices of giving and acting for reasons.

iv. Imagination, in its more usual creative, generative, possibility-generating guise, was more generally proposed as the crucial play- and testing-ground of the free agent, in which possible futures are conjured and worked out. More speculatively, it was hypothesised that giving imagination its dues within an empirically informed, less rationalistic account of free agency might help incompatibilism locate a form of indeterminism that is not only freedom friendly but recognisably freedom enhancing.

v. Language, and linguistically and symbolically mediated concepts and reasoning, are of course critical to a full account of what makes human agency distinctive. But even here, I have argued, we can look at reasons for the significance of these capacities other than that of enabling a (linguistically facilitated) capacity to constrain the will based on the demands of reason – traditionally something especially prized in various structural accounts of the will. Instead, we can find here fruitful material for identifying a variety of degrees of freedom that become accessible to human agents who have successfully entered a literate cognitive community. The creative power of a symbolically-endowed imagination, the ability to transcend individual capacity limitations via symbolic technologies, and the mnemonic and epistemic bootstrapping effects enabled by the externalising and recording of knowledge and skill in cognitive communities, are just three of the promising degrees of freedom that I have highlighted for further exploration.

The significance of the contributions listed under (i) to (v) is not exhausted, however, by their acting as a brake on the tendency to imagine or mistake ourselves as/ for HHAs, nor by the promise they hold for shaping fresh, positive, empirically-informed accounts of the nature and constraints of free agency – these conclusions also help bolster the response made in this thesis to the sceptical empirical case that we are, or might be, AAs. *Contra* the sceptical views of conscious will and self-understanding espoused by Wegner and Bargh, the discussions and data regarding self-insight and verbal overshadowing in Chapter 9 suggest the possibility of a more positive and constructive exercise of understanding our *real* capacity for conscious reflective deliberation without either exaggerating or underestimating its limits and constraints.

Contra the picture of emotional and other 'automatic' reactions offered in (especially) the automaticity literature, emotion is not something to be seen as in conflict with our attempts at conscious self-governance. Emotion can be troublesome, and it can operate before and without us knowing it. But the efficient operation of emotional systems is, on the whole, an integral adaptive, reliable and trustworthy aspect of our conscious and deliberative engagements with the world.

Similarly, the idea that automated responses are somehow in conflict with, and threatening to, any claims for effective conscious self-governance is exposed as a most peculiar assumption. On the views surveyed under (iii), our capacity to automatise complex, skilled behavioural sequences is one of the key virtues of our distinctive capacity for conscious agency and self-governance. It is precisely against a backdrop of skill acquisition through progressive refinement and automation that the real power of multifocal attention and monitoring can be realised – now, or in general, focussing on the 'big picture', but equally able to focus in on a detail for fine-tuning. These capacities have added degrees of freedom to our agency, not alienated us from ourselves.

Such detailed observations of our agency at work in the real world also help expose the severe shortcomings of Wegner's model of apparent mental causation, as they do the flaws in Libet's underlying assumptions about the 'mechanics' of consciously chosen voluntary movements. These models are hopelessly simplistic, decontextualised, and excessively linear. They offer no sensible way in which they could be mapped onto the complex dynamics of

conscious agency as manifested in the temporally and spatially extended framework of intermediate-term governance. Furthermore, none of the sceptical arguments do justice to the complex, interactive and distributed 'looping' effects that the likes of Gallagher have highlighted as such a distinctive, powerful and liberating marker of human conscious agency.

Returning to the positive project, I have proposed that a recognition, and empirical and philosophical exploration, of the nature and roles of imagination and various externalised aspects of mind should provide a focus for future work that, in the spirit of the perspective on free agency taken, explicated and defended in this thesis, seeks to situate, explicate and defend our claims of freedom within the parameters of an interdisciplinary study of the mind.

In summary of Part II, then, I have argued for an integrative response to certain pressing empirical challenges to the idea that we might be consciously self-governed agents worthy of being described as free agents. Through a combination of philosophically- and empirically-grounded arguments, I have attempted to dismantle these cases for our being AAs. I have used the same strategy to argue against any tendency to unpack claims about human agency in general, or free agency in particular, in ways that require us to be (or become) HHAs. In the process of pursuing these two goals, I have both articulated and advertised what I take to be promising lines for further exploration and defence of important and distinctive of aspects of our agency that can ground claims that we are free agents outside of the traditional philosophical framework of the compatibilist-incompatibilist free will debate. It is, I think, the future pursuit of research along these lines, fertilised by greater dialogue between philosophy, psychology, and the interdisciplinary sciences of the mind, and unfettered by a deep and stubborn impasse, that can promise to deliver an empirically viable and philosophically sound account of free agency.

(Re-)Addressing Tradition[249]

The discussions and arguments of Part II of this thesis take as their point of departure an attempt to reframe the debate about free agency, as laid out in Chapter 5. For those not convinced by the arguments of that chapter, and by the arguments of Part I more generally, the balance of the thesis can at least be read as an invitation to begin exploring where we might get in making a case for free agency by shifting focus away from the controversies and

[249] This re-addressing of positions and issues in the traditional debate was requested by one of the examiners of the thesis.

major points of contention within the traditional debate about free will. There is, I think, more to be done in this direction before attempting to make a more detailed mapping of issues and positions within and between the traditional debate and my proposed alternative problem space sketched in Chapter 5. Any premature attempts at such a mapping carry the risk of distorting potentially novel ideas, positions and problems by forcing a fit to the much more established, well-chartered terrain of tradition. There is much potential for misconstruing the value of the projects in this thesis, both embarked on and advertised for future work, by simply labelling them, embracing them, and/or dismissing them as one more variety of compatibilism/ incompatibilism/ libertarianism/ scepticism.

Nevertheless, the weight of conceptual and terminological usage favours tradition, and so I will close by saying something more about how the position I have been developing in the thesis, and in particular the discussions of Part II that draw on empirical work in psychology and allied fields, relates to and advances the traditional philosophical discussion of free will and agency.

Given my arguments, assessments and conclusions of Part I, my own position would be a variety of what I would call *free will incompatibilism*. It is an incompatibilist position in so far as the global project of compatibilism is rejected as a failure. It is a variety of free will incompatibilism because it seeks to defend a claim that we are free agents in a philosophically and psychologically significant sense. It is thus to be distinguished from sceptical incompatibilist positions that would take the failure of compatibilism and the apparent weaknesses of traditional libertarianism to imply insurmountable difficulties for any would-be defender of free agency. Finally, I call the position free will incompatibilism so as to distinguish it from these traditional varieties of libertarianism, given my negative assessments of Chapter 2.

Certainly, my position does not yet comprise a fleshed out account of free agency, let alone a comprehensive theory of human agency. But as a non-traditional variety of non-sceptical incompatibilism, it is a position that endeavours to draw on the wisdom of tradition while avoiding its problems and excesses. Here is an outline of how I see it doing so.

With the compatibilist, I am keen to place emphasis on (broadly speaking) the role of an agent's *character* in guiding deliberation, choice and action. With Kane, I think that the

threat of determinism does call for the presence of some regress-halting moments of indeterminism in character formation and self shaping. But *contra* Kane, I do not propose to locate this indeterminism in special moments of rational choice characterised by plural voluntary control (i.e. aspects of Kane's SFAs that make them too much like compatibilist rational choice). Instead, I have hypothesised that indeterministic moments of self shaping might sensibly be located within processes of imaginative generativity in which our future possibilities are imagined.

How might this advance the traditional debate? Because I am not committed to an incompatibilist analysis of choice situations – either in general, or in the case of occasional special instances such as Kane's SFAs – I am able to embrace much that seems sensible and valuable in what compatibilists have to say at a local level about rational deliberation and choice, as well as the conditional analysis of alternative possibilities in these situations. I am also able to take on a roughly compatibilist view of what seemed like a descriptively and normatively appealing picture of how we progressively mature and become free agents over time – the picture of emerging agency sketched by Dilman, as well as by Mele, as outlined in Chapter 3 – that seeks a balance between an acceptance and endorsement of what we find ourselves to be, on one hand, and processes of critical self-evaluation and gradual self shaping on the other hand.

At the same time, because the overarching compatibilist project is best abandoned by my lights, there is no longer a need to ratchet up the demands of rationality, self-awareness, self-reflection, self-control, etc. to unrealistic levels (given the evidence and arguments of Chapters 6 and 9) in order to persuade compatibilist critics that such 'hyper-agents' would qualify as free even in a deterministic universe. Instead of being evaluated against the criteria of the traditional debate, then, what is worth saving from compatibilist accounts will be evaluated according to its descriptive and normative 'goodness-of-fit' with the psychologies of real human agents who are nevertheless realistic candidates for qualifying as free agents.

While this much good sense in compatibilist accounts can be saved, my position remains a fundamentally incompatibilist one. As wished for by traditional libertarians, I share the view that agents need to be freed from the causal straightjacket of global determinism by positing appropriate moments of indeterminism in their histories. But I try to do so without positing either special or regular moments of *volitional* indeterminism that might threaten the agent's

claims of ownership and control over important free choices. My hypothesis about imagination and imaginative generativity discussed in Chapter 10 thus represents an attempt to balance the joint desiderata of positing agency-enhancing, self-shaping moments of indeterminism while not undermining general claims of agency, ownership and control over choice and action.

My comments so far highlight implications of the discussions and arguments of Part I, Chapter 5 and Chapter 10 for the traditional debate. Much of Part II, however, is occupied with the task of developing a more empirically and psychologically based analysis of agency. It is this part of the project that is both more difficult to map onto the traditional debate, as well as being the aspect I am most reluctant to prematurely cast in the mould of the concepts, categories and positions that mark the tradition's well-trodden topography. However, I will end by reflecting on how I see this empirically-informed work as taking the debate forward, and how I see the contribution it makes to the development of my own position on free agency.

An important dimension of what I have attempted in Part II of the thesis is to demonstrate that there are pressing questions and puzzles about agency that require attention if we are to develop and defend a comprehensive, coherent and empirically-plausible account of human free agency. These questions, initially posed against the backdrop provided by the image of the Agent Automaton, are ones (i) that do not necessarily flow from positions, assumptions and key points of contention within the traditional debate; (ii) that require attention from all those interested in defending free agency, whatever their traditional allegiances, and irrespective of whether or not they are persuaded by my critical arguments in Part I; and (iii) whose posing and answering arguably requires (and thus contributes) a fresh, empirically-informed and nuanced perspective on human agency that goes beyond much of the abstract, non-occurrent belief-desire psychology that runs through so much traditional philosophical theorising about agency.

The questions, challenges and puzzles raised in articulating a 'new' form of sceptical threat to shape the debate (Chapter 6), as well as responding to this threat while avoiding the potentially problematic abstractions and excesses of tradition (Chapters 7 to 10), consistently highlight the importance of a number of themes, characteristics and processes for our understanding of agency: consciousness; occurrent mental activity; forms, systems and

processes of control; the nature, power and limits of self-awareness; and the nature of limits of (embodied, real-world) human rationality.

My contention has been that, to the extent that the traditional debate does not, and in general does not seek to address these puzzles and questions, we are missing important parts of a comprehensive account of human free agency that is equipped to respond to sceptical claims about agency that are not fundamentally grounded in considerations and concerns about determinism. My efforts to both frame and respond to some of the most pressing challenges within the alternative problem space sketched in Chapter 5 thus not only represent an attempt to make progress towards completing such a comprehensive account. The exercise itself aimed to demonstrate the potential value of abandoning (or at least temporarily setting aside) the traditional debate and its problem space of issues and positions, in order to take and develop fresh perspectives on human agency and freedom that are not primarily motivated or constrained by that tradition. Those unwilling to work outside the bounds of the tradition are nevertheless invited to take on board the arguments and insights I have offered, adjusting their accounts of agency as and where my analyses and the relevant empirical data call for this.[250]

What about the development of my own position over the course of Part II? There are two overarching ideas that are at work in my emerging account of free agency. The first involves taking up O'Connor's challenge of giving adequate attention to consciousness within an account of free will. The second involves the idea that we lay claim to being free agents by accruing and exploiting various *degrees of freedom*, rather than by just satisfying one set of (compatibilist or libertarian) conditions for having exercised our free will in a given instance. I will expand on each of these ideas in turn.

[250] I imagine, in other words, that various compatibilists, and libertarians such as Kane (but probably not agent-causal libertarians), might read significant portions of Part II and be willing to take much of what I have discussed and argued on board, adjusting their accounts of agency where necessary. While I clearly think there is greater value to be gained by setting the traditional debate aside, it is no threat to the value and import of my claims and arguments if they can be taken on board by certain adherents to the tradition. Nor does the potential compatibility of my claims with, say, a given variety of compatibilism mean that I am (in the end) just another compatibilist. What I reject in setting aside the traditional debate is the idea that the only important claims and insights about free agency are ones that further the interests of one or other traditional camp within the debate at the expense of one or more of the others. The threat of the Agent Automaton, and the skeptical arguments of Chapter 6 in general, show that we do not need to be worrying about determinism in order to worry about, explicate and defend claims of free agency.

I agree with O'Connor that the traditional debate in general, and compatibilist accounts in particular, do not do justice to the role and significance of consciousness in our distinctive form of agency; and the sceptical challenges posed in Chapter 6 help highlight the urgency of filling in the resulting gaps. From Chapter 7 onwards, I take up this challenge in at least four ways: (i) putting forward a number of conceptual and philosophical clarifications that move the discussion away from the crude models of conscious agency apparently at work within the sceptical arguments of Chapter 6; (ii) carefully delimiting the likely nature and role of distinctively conscious modes of agency and control by foregrounding the multiplicity of modes of systemic organisation and control we should expect to find at work within a complex biological system like a human being; (iii) carefully delimiting the likely nature, role and extent of reflexive conscious self awareness, given the apparent problems that arise from positing (and/or prescribing) an excessively abstract, introspective, and coldly rational deliberative position, unrealistically distanced from the complex dynamics of embodied, emotional, world-directed agency; and (iv) exploring and highlighting aspects of human agency – imagination and externalised aspects of mind – that are not only distinctive and freedom enhancing, but that also unequivocally and ineliminably draw on the distinctive character and power of human conscious processing, control and self-governance over intermediate-term timeframes.

Looking ahead to the further work that needs to be done in pursuit of a theory of free agency, it is this last idea of processing, control and self-governance over the intermediate term that promises to form the backbone of a positive defence of our free agency in the face of sceptical challenges, both empirical and traditional. In recognising that the most distinctive and powerful roles for conscious agency lie at the level of intermediate-term control, guidance and self-governance, much of the apparent threat of our being Agent Automatons is immediately diffused. Along with Dennett, Gallagher and others[251], we can see that what is most distinctive about human agency is not so much something about knife-edge choice situations and micro-initiation and management of neural commands and muscle movements. Intermediate-term conscious control and governance is primarily focussed on larger scale and longer term intentional goals and projects that are sustained in the face of a range interfering factors, of both internal and external varieties.

[251] And *contra* agent-causal libertarians.

The notion of accruing and exploiting degrees of freedom is explicitly mentioned in Chapter 10, but is implicitly at work in the preceding chapters that advance our understanding of the nature of conscious agency and will. In essence, the idea is that abandoning the main positive traditional projects of libertarianism and compatibilism leaves open the possibility of grounding free agency in *multiple* features of human agency, rather than in one tightly-knit set of sufficient (or necessary and sufficient) conditions for free will that seeks to address one particular set of concerns (making indeterministic volition plausible; making free will in a deterministic universe plausible). This is a point that can be highlighted by again returning to my hypothesis regarding indeterministic processes of imaginative generativity. It is nowhere a part of my thinking about these processes that they and they alone could ground claims to our being free agents. My hypothesis is that this would be a highly plausible and appealing place in which to locate a regress-halting form of indeterminism that would block the kinds of backwards-stretching causal chains that incompatibilists see as ultimately undermining of freedom. But I do not think that these processes could ground claims of Ultimate Responsibility (*a la* Kane); nor do I think that they could provide anything like the whole story about human free agency. They are supposed to provide a buffer against freedom-threatening determinism; but they do so *as part of* a more complex set of processes and capabilities displayed by most normal adult human agents. Included amongst these are our capacities for deploying kinematic imagination, and engaging and exploiting various externalised aspects of mind that help us (amongst other things) transcend our otherwise inherent physical, biological and psychological capacity limitations. But also included would be many of the capabilities rightly valued by varieties of compatibilism, including processes associated with reflective endorsement, gradual self-shaping, and giving and demanding reasons for action of ourselves and others. The task of developing and completing a comprehensive theory of free agency will require further articulation of how the relevant processes and capabilities fit together to confer a meaningful and defensible form of free agency that is, ultimately, incompatibilist.

Is this just thinly disguised compatibilism? Or a compatibilism with an aesthetically-pleasing sprinkling of indeterminism? No – the elements of compatibilist accounts of agency that survive must do so because they correctly characterise human agency, and because these elements contribute important degrees of freedom. Recognising these virtues within certain extant compatibilist theories in no way overturns the judgement that the grand or overarching compatibilist project is a failure. Moreover, my commitment to finding the appropriate

place/s for indeterminism within a positive incompatibilist account remains firm, even if the current project only goes so far as to generate a speculative working hypothesis as to how this should be done.

But is a 'promissory' commitment to an indeterministic theory of free agency tenable, given the admittedly poor track record of libertarian attempts to put indeterminism to work in service of freedom instead of undermining it? I think the commitment is tenable, not only because I think that there is real value to be gained by taking my hypothesis about imagination further, but also because I think that Kane is right to highlight (as we saw in Chapter 2) a prejudice we tend to bring to our thinking about non-deterministic causal processes, such that 'indeterministically caused' is too easily equated with 'uncaused'. Correctly diagnosing and combating this prejudice is, I think, part of the larger project I mentioned in Chapter 5 of expunging reductionistic physicalist thinking from the philosophy of mind and psychology, and is thus again a task that lies beyond the scope of the current project. But as a rough diagnosis, at the heart of the matter lies the tendency to dichotomise deterministic causation and chance, as if every causal process that is less than deterministic becomes a *mere* matter of chance. This is a false dichotomy, however, and it is as potentially misleading as the simplistic identification of determinism with fatalism. My contention is that a proper evaluation of my proposed brand of incompatibilism must await something of a change in custom and practice in how we think about indeterministic or 'chancy'[252] causation[253].

This is in no way an attempt to sidestep the issues at stake. I acknowledge the gaps that remain, and the work that needs to be done. But I take heart that what might strike some as weird and implausible need not turn out to be so, even if some of the weirdness were to remain. Jerome Kagan, commenting on the difficulties inherent in measuring emotion and feeling given the subtle complexities involved, hints amusingly at just the kind of prejudice that is often shown by science towards things mental despite some of the oddities of the unambiguously physical:

> If the rational deductions of physicists... require us to believe that a stream of photons directed at a screen with two slits passes through both openings simultaneously (medieval Christian scholars

[252] To borrow a label from Hugh Mellor.
[253] Libertarian incompatibilists might stand to gain something from such a change in thinking about indeterministic causal processes too; but in my assessment, their problems did not lie in positing indeterminism *per se* so much as in *where* they want to insert it – namely in moments of volition.

> speculated that an angel could be in two places simultaneously), and a temporal interval too short to imagine was sufficient to allow the marble-sized bundle of energy that existed before the big bang to expand to a universe measured in billions of light-years, it may be time to honor the legitimacy of the emotion English calls “confusion,” even if this state is difficult to measure with the precision Lord Kelvin demanded. (Kagan, 2007, p108)

One could easily (if less elegantly) add honouring the legitimacy of human free agency, conceived along incompatibilist but non-libertarian lines, incorporating suitable freedom-enhancing indeterministic processes. The traditional debate about free will has not exhausted the possibilities for exploring how to make such an account work; this thesis has sought to begin this work, and outline avenues along which it may proceed in the future.

References

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

Allison, H.E. (1990). *Kant's theory of freedom*. Cambridge: Cambridge University Press.

Baars, B.J. (1994). A Global Workspace theory of conscious experience. In A. Revonsuo & M Kamppinen (eds.) *Consciousness in philosophy and cognitive neuroscience.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Baars, B.J. (1997). Treating consciousness as an empirical variable: The contrastive analysis approach. In N. Block, O. Flanagan, & G. Guzeldere (eds.) *The Nature of Consciousness: Philosophical Controversies*. Cambridge, MA: MIT Press.

Baer, J., Kaufman, J.C. & Baumeister, R.F. (2008). *Are we free: psychology and free will*. New York: Oxford University Press.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230-244.

Bargh, J. A., Raymond, P., Pryor, J., & Strack, F. (1995). Attractiveness of the underlying: An automatic power-sex association and its consequences for sexual harassment and aggression. *Journal of Personality and Social Psychology*, 68, 768-781.

Bargh, J.A. (2008). Free will is un-natural. In J. Baer, J.C. Kaufman & R.F. Baumeister (eds.). *Are we free: psychology and free will*. New York: Oxford University Press.

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

Baron-Cohen, S. (1995). Mindblindness: an essay on autism and theory of mind. Cambridge, MA.: MIT Press.

Bechara, A., Damasio, A.R., Damasio, H. & Anderson, S.W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7-15.

Bechara, A., Damasio, H. & Damasio, A.R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10, 295-307.

Bechara, A. & Damasio, A.R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Games and Economic Behavior*, 52, 336–372.

Blackburn, S. (1998). *Ruling passions: a theory of practical reasoning*. Oxford: Clarendon Press

Blackburn, S. (1999). *Think: a compelling introduction to philosophy*. Oxford: Oxford University Press.

Blackmore, S. (1999). *The meme machine*. Oxford: Oxford University Press.

Blackmore, S. (2005). *Conversations on consciousness*. Oxford: Oxford University Press.

Brook, A. and Ross, D. (eds.) (2002). *Daniel Dennett*. Cambridge: Cambridge University Press.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4, 73-121.

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32, 121–182.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Cartwright, N. (1999). *The dappled world: a study of the boundaries of science*. Cambridge: Cambridge University Press.

Chen, M. & Bargh, J.A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, 33, 541—560.

Chin, J.M. & Schooler, J.W. (2008). Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. *European Journal of Cognitive Psychology*, 20, 396-413.

Chomsky, N. (1959). Review of B F Skinner's *Verbal behaviour*. *Language*, 35, 26-58.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 58, 7–19.

Clark, A. (1989). *Microcognition*. Cambridge, MA: MIT Press.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again.* Cambridge, MA, MIT Press.

Clark, A. (2008). Supersizing the mind: embodiment, action, and cognitive extension. Oxford: Oxford University Press.

Clarke, R. (1999). Free choice, effort, and wanting more. *Philosophical Explorations*, 2, 20-41.

Clarke, R. (2003a). Toward a credible agent-causal account of free will. *Noûs*, 27,191-203.

Clarke, R. (2003b). *Libertarian accounts of free will*. Oxford: Oxford University Press.

Clarke, R. (2005). Agent causation and the problem of luck. *Pacific Philosophical Quarterly*, 86, 408–421.

Crane, T. & Mellor, D.H. (1995). *Postscript* to 'There is no question of physicalism.' In P.K. Moser & J.D. Trout (eds.) *Contemporary metaphysics: a reader*. London: Routledge.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary Psychology and the generation of culture*. New York: Oxford University Press.

Damasio, A.R. (1994). *Descartes' error: emotion, reason and the human brain*. London: Papermac/Macmillan.

Damasio, A.R. (1999). *The feeling of what happens: body, emotion and the making of consciousness*. London: Vintage.

Davidson, D. (1963). Actions, reasons, and causes. Reprinted in D. Davidson (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1970). Mental events. Reprinted in D. Davidson (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1974). Psychology as philosophy. Reprinted in D. Davidson (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Davidson, D. (1987). Knowing one's own mind. Reprinted in D. Davidson (2001). *Subjective, intersubjective, objective*. Oxford: Clarendon Press.

Deacon, T.W. (1997). *The symbolic species: the co-evolution of language and the brain*. New York: Norton.

Dennett, D.C. (1984). *Elbow Room*. Cambridge, Mass.: MIT Press.

Dennett, D.C. (1987), *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D.C. (1991). *Consciousness Explained*. London: Penguin.

Dennett, D.C. (1992). Review of Varela et al. & Edelman. *New Scientist*, 13 June, 48-9.

Dennett, D.C. (1995). *Darwin's dangerous idea: evolution and the meanings of life*. London: Penguin.

Dennett, D.C. (1996). Cow-sharks, magnets, and Swampman. *Mind and Language*, 11, 76–77.

Dennett, D.C. (2003). *Freedom evolves*. London: Penguin.

Dennett, D.C. (2005). Natural Freedom. *Metaphilosophy*, 36, 449-459.

Dennett, Daniel C. (13 June 1992). Review of Varela et al. & Edelman. *New Scientist.* 48-49

Dias, M.G. & Harris, P.L. (1990). The influence of the imagination on reasoning by young children. *British Journal of Developmental Psychology*, 8, 305-318.

Dilman, I. (1999). *Free will*. London: Routledge.

Donald, M. (2001). *A mind so rare: the evolution of human consciousness*. New York: Norton.

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

Dunn, B.D., Dalgleish, T. & Lawrence, A.D. (2006). The somatic marker hypothesis: a critical evaluation. *Neuroscience and Biobehavioral Reviews*, 30, 239–271.

Dupré, J. (1993). *The disorder of things: metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Edelman, G.M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.

Edelman, G.M. (1988). *Topobiology: an introduction to molecular embryology*. New York: Basic Books.

Edelman, G.M. (1989). *The remembered present*. New York: Basic Books.

Edelman, G.M. (1992). *Bright air, brilliant fire*. New York: Basic Books.

Edelman, G.M. (2004). *Wider than the sky: the phenomenal gift of consciousness*. London: Allen Lane.

Edelman, G.M. and Tononi G. (2000). *Consciousness: How Matter Becomes Imagination*. London: Allen Lane.

Ekstrom, L.W. (2002). Libertarianism and Frankfurt-style cases. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

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

Fischer, J.M. (1994). *The metaphysics of free will: an essay on control.* Oxford: Blackwell.

Fisher, J.M. (2002). Frankfurt-type examples and semi-compatibilism. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

Fodor, J. A. (1992). The big idea: Can there be a science of mind? *Times Literary Supplement*, July, 5-7.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66, 829-839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68, 5-20. Reprinted in G. Watson (ed.) (1982). *Free Will*. Oxford: Oxford University Press.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Clarendon Press.

Gazzaniga, M. & Gallagher, S. (1998). The Neuronal Platonist: Michael Gazzaniga in conversation with Shaun Gallagher. *Journal of Consciousness Studies*, 5, 706–717.

Gollwitzer, P.M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist*, 54, 493-503.

Guzeldere, G. (1995a). Consciousness: What it is, how to study it, what to learn from its history. *Journal of Consciousness Studies*, 2, 30-51.

Guzeldere, G. (1995b). Problems of consciousness: a perspective on contemporary issues, current debate. *Journal of Consciousness Studies*, 2, 112-143.

Haji, I. (1999). Indeterminism and Frankfurt-type examples. *Philosophical Explorations*, 2, 42-58.

Haji, I. (2002). Compatibilist views of freedom and responsibility. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

Harris, P.L. (2000). *The work of the imagination*. Oxford: Blackwell.

Harris, P.L. (2001). The veridicality assumption. *Mind & Language*, 16, 247-262.

Harris, P.L., German, T. & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61, 233-259.

Holland, J.H. (1998).*Emergence: from chaos to order*. Oxford: Oxford University Press.

Honderich, T. (2002). Determinism as true, compatibilism and incompatibilism as false, and the real problem. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

Horgan, J. (1996). *The end of science: facing the limits of knowledge in the twilight of the scientific age*. London: Abacus.

Ismael, J.T. (2007). *The situated self*. Oxford: Oxford University Press.

James, W. (1897/2006). *The Will to Believe and Other Essays in Popular Philosophy*. New York: Cosimo.

Jarrold, C. (2003). A review of research into pretend play in autism. *Autism*, 7, 379-390.

Johnson, S. (2001). *Emergence*. London: Penguin.

Kagan, J. (2007). *What is Emotion?* New Haven: Yale University Press

Kane, R.H (ed.) (2002c). *The Oxford handbook of free will*. Oxford: Oxford University Press.

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

Kane, R.H. (1999a). Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy*, 96, 217-240.

Kane, R.H. (1999b). On free will, responsibility and indeterminism: responses to Clarke, Haji, and Mele. *Philosophical Explorations*, 2, 105-121.

Kane, R.H. (2002a). Free will: new directions for an ancient problem. In R. Kane (ed.) *Free Will*. Oxford: Blackwell.

Kane, R.H. (ed.) (2002b). *Free Will*. Oxford: Blackwell.

Kant, I. (1781/1787/2007). *Critique of Pure Reason*. London: Penguin.

Kant, I. (1788/2010). *Critique of Practical Reason*. New York: Classic Books.

Kirsch, I. & Lynn, S.J. (1999). Automaticity in clinical psychology. *American Psychologist*, 54, 504-515.

Korsgaard, C.M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

LeDoux, J.E. (2002). *Synaptic self: how our brains become who we are*. New York: Penguin.

Leevers, H.J. & Harris, P.L. (1998). Drawing impossible entities: A measure of the imagination in children with autism, children with learning disabilities, and normal 4-year-olds. *Journal of Child Psychology and Psychiatry*, 39, 399-410.

Libet, B. (1985). Unconscious cerebral initiative and the role of the conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529–66.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, 6, 47-57.

Libet, B. (2004). *Mind time: the temporal factor in consciousness*. Cambridge, MA: Harvard University Press.

Libet, B., Freeman, A. & Sutherland, K. (1999). *The volitional brain: towards a neuroscience of free will.* Thorverton: Imprint Academic. Also published as Issues 8 and 9 of the *Journal of Consciousness Studies*, Volume 6, 1999.

Libet, B., Gleason, C.A., Wright, E.W. & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*, 106, 623–642.

Lipton, P. (2004). Genetic and generic determinism: a new threat to free will? In D.A. Rees & S.P.R. Rose (eds.) *The new brain sciences: perils and prospects*. Cambridge: Cambridge University Press.

Locke, J. (1690/1975). *An Essay Concerning Human Understanding*. Oxford: Clarendon Press.

Marcel, A.J. (1988). Phenomenal experience and functionalism. In A.J. Marcel & E. Bisiach (eds.) *Consciousness in Contemporary Science*. Oxford: Clarendon Press.

McCrone, J. (1999). A bifold model of free will. *Journal of Consciousness Studies*, 6, 241–259.

McGinn, C. (2004). *Mindsight: image, dream, meaning*. Cambridge, MA: Harvard University Press.

McKay, T. & Johnson, D. (1996). A reconsideration of an argument against compatibilism. *Philosophical Topics*, 24, 113-122.

Melcher, J.M. & Schooler, J.W. (1996). The misremembrance of wines past: verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, 35, 231–245.

Mele, A.R. (1992). *Springs of action*. New York: Oxford University Press.

Mele, A.R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.

Mele, A.R. (1999). Kane, luck, and the significance of free will. *Philosophical Explorations*, 2, 96-104.

Mele, A.R. (2005). Dennett on freedom. *Metaphilosophy*, 36, 414-426.

Mele, A.R. (2006). *Free Will and Luck*. Oxford: Oxford University Press.

Mele, A.R. (2008). Psychology and free will: a commentary. In J. Baer, J.C. Kaufman & R.F. Baumeister (eds.). *Are we free: psychology and free will*. New York: Oxford University Press.

Meltzoff, A.N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838-850.

Nagel, T. (1986). *The View from Nowhere*. New York: Oxford University Press.

Nahmias, E., Morris, S., Nadelhoffer, T. & Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies*, 11, 162–179.

Nichols, S. (2007). The rise of compatibilism:A case study in the quantitative history of philosophy. *Midwest Studies in Philosophy*, 31, 260-270.

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

Nørretranders, T. (1998). *The user illusion: cutting consciousness down to size*. London: Penguin.

Nozick, R. (1981). *Philosophical explanations*. Oxford: Clarendon Press.

O'Connor, T. (1995a). Agent causation. In T. O'Connor (ed.) *Agents, causes, and events: essays on indeterminism and free will*. New York: Oxford University Press.

O'Connor, T. (ed.) (1995b). *Agents, causes, and events: essays on indeterminism and free will*. New York: Oxford University Press.

O'Connor, T. (2000). *Persons and causes: the metaphysics of free will*. New York: Oxford University Press.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.

Reeke, G.N., & Edelman, G.M. (1984). Selective networks and recognition automata. *Annals of the New York Academy of Sciences*, 426, 181-201.

Resnick, M. (1994). *Turtles, termites, and traffic jams: explorations in massively parallel microworlds*. Cambridge, MA: MIT Press.

Resnick, M. (1996). Beyond the centralized mindset. *Journal of the Learning Sciences*, 5, 1-22.

Rose, S.P.R. (1997). *Lifelines: life beyond the gene*. New York: Oxford University Press.

Russell, P. (2002). Pessimists, Pollyannas, and the New Compatibilism. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

Saver, J.L. & Damasio, A.R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, 29, 1241-1249.

Schooler, J.W. & Engstler-Schooler, T.Y. (1990). Verbal Overshadowing of Visual Memories: Some Things Are better Left Unsaid. *Cognitive Psychology*, 22, 36-71.

Schroeter, F. (2004). Endorsement and Autonomous Agency. *Philosophy and Phenomenological Research*, 69, 633-659.

Searle, J.R. (2001a). Free will as a problem in neurobiology. *Philosophy*, 76, 491-514.

Searle, J.R. (2007). *Freedom and neurobiology: reflections on free will, language, and political power*. New York: Columbia University Press.

Searle, J.R. (1997). *The mystery of consciousness*. London: Granta.

Searle, J.R. (2001b). *Rationality in action*. Cambridge, MA: MIT Press.

Smilansky, S. (2000). *Free will and illusion*. Oxford: Clarendon Press.

Spencer, S.J., Fein, S., Wolfe, C.T., Fong, C. & Dunn, M.A. (1998). Automatic activation of stereotypes: the role of self-image threat. *Personality & Social Psychology Bulletin*, 24, 1139-1152.

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, 75, 5–24.

Strawson, G. (2004). Real intentionality. *Phenomenology and the Cognitive Sciences*, 3, 287–313.

Thelen, E. and L. Smith. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies*, 27, 185-199. Reprinted in G. Watson (ed.) (1982). *Free Will*. Oxford: Oxford University Press.

van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon Press.

van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives*, 14, 1-19.

Want, S.C. & Harris, P.L. (2001). Learning from other people's mistakes: Causal understanding in learning to use a tool. *Child Development*, 72, 431–443.

Watson, G. (1975). Free agency. *Journal of Philosophy*, 72, 205-220. Reprinted in G. Watson (ed.) (1982). *Free Will*. Oxford: Oxford University Press.

Wegner, D.M. & Wheatley, T. (1999). Apparent mental causation: sources of the experience of will. *American Psychologist*, 54, 480-492.

Wegner, D.M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Widerker, D. (2002). Responsibility and Frankfurt-style examples. In R.H Kane (ed.) *The Oxford handbook of free will*. Oxford: Oxford University Press.

Williams, B. (1976). Persons, character and morality. Reprinted in B. Williams (1981). *Moral Luck*. Cambridge: Cambridge University Press.

Wilson, T.D. & Dunn, E.W. (2004). Self-knowledge: its limits, value, and potential for improvement. *Annual Review of Psychology*, 55, 493–518.

Wilson, T.D. & Schooler, J.W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60, 181–92.

Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.