THE NATURE OF FREE WILL

DAVID THOMAS WILSON

M.Sc., LL.B., B.A. (Hons.)

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Department of Philosophy
Macquarie University

February, 2006

TABLE OF CONTENTS

SUMMARY

There is more than one problem of free will. Many philosophers approach the free will question seeking a foundation for systems of ethics or a justification for societal practices of praise and blame. Important as those issues are, they are not my quest. Rather, I address the metaphysical question of how to accommodate free will within the natural world. I conclude that the natural world is not identical with the physical world, and that it must contain entities or influences that are not physical in any current sense of that word.

The view I defend is a dualist view, but I claim dualism could be avoided only by accepting premises that are less acceptable. Humans appear to exercise choices that affect the physical history of the world, and if that appearance were merely an illusion, there is no plausible explanation for the evolution of consciousness in our species.

There are many senses in which agents may be described free to act in some way. Through a linguistic analysis of the modal verbs CAN and COULD, I distinguish one underlying concept that I define as "metaphysical free will" in Chapter 2.

Many philosophers contend that free will is compatible with causal closure of the physical domain. In Chapter 3, I argue that "metaphysical free will" is not so compatible, and that compatibilist arguments involve a different concept of free will.

The reality of metaphysical free will and its incompatibility with physical causal closure entail that the physical domain is not causally closed. In chapter 4, I explore why many philosophers endorse causal closure, and reject arguments in its favour.

In Chapter 5, I argue that metaphysical free will can be reconciled with physics though not subsumed within it, and in the final chapter, I describe some current projects attempting to explain consciousness or free will in terms of quantum theories of physics. I suggest that any theory successfully accounting for free will will extend beyond physics, but fall within a wider view of "Nature".

I certify that the following thesis is entirely my own original work and has not been submitted for a higher degree to any other University or educational institution. All sources of information used in the thesis have been indicated, and due acknowledgement has been given to the work of others.

Signed:

David Thomas Wilson

Date: February, 2006

ACKNOWLEDGEMENTS

I express my sincere thanks to the staff and to my fellow students in the Department of Philosophy at Macquarie University, for making me feel welcome, and for providing such a congenial working environment.

In particular, I thank my supervisor, Peter Menzies, for his encouragement and guidance in this project. Though I know he does not share some of the conclusions I am putting forward, he has always been ready to consider them with an open mind, while making me aware of contrary opinions. Our discussions on the points where we continue to disagree have helped me to clarify my views, by forcing me to defend them.

I thank my daughter, Jenny Duke-Yonge, for constructive criticism, for expert proofreading, and especially for kindling my interest in philosophy. I thank my son, Hugh Wilson, for his expert advice on physics, and for his critical reading of the final three chapters.

Lastly, I thank my wife, Mary, for her patience and for the encouragement she has given me to see this project through to completion.

CHAPTER 1

FREE WILL AND NATURE

1.1 The Nature of Free Will

1.1.1 Goals and Assumptions

This thesis investigates "the nature of free will", in two senses of the word "nature". Firstly, I address the "nature" of free will in the sense that I seek to identify the intrinsic or essential character of what we call free will. In fact, several concepts go by the name of "free will", but I shall identify and distinguish one fundamental concept of free will that I claim plays an important role in the lives of human beings and probably of other intelligent living creatures. That concept of free will is reflected in human languages, and through an analysis of part of one language – English – I shall specify the "nature" of the free will so conceptualised, and show how that concept may be distinguished from other concepts that in different contexts can be described by the phrase "free will".

The second sense in which I seek to elucidate the "nature" of free will is to consider what must be concluded about "Nature" – the entire natural realm –, if free will in the sense that interests me is to be acknowledged as a part of it. I argue that Nature in that sense must include properties or entities beyond those encompassed by the physical sciences, according to any normal understanding of the word "physical".

The problem at the heart of this thesis is an old one. How can the widely-held belief that human agents can and do sometimes make free choices influencing the

world around them be reconciled with the natural order of things, as that order seems to be revealed by empirical science? I do not claim to have found a complete solution to that problem. It is a problem about which philosophers have argued for at least two and a half millennia without reaching a consensus. My first aim in this thesis is to show that the problem is a genuine one, and to separate it from other more tractable problems which I believe are independent of it. My further aims are to show that some of the solutions – or approaches to solutions – that currently find favour within the philosophical community are untenable, and by eliminating those approaches, to indicate the direction in which an answer to the age-old question is more likely to lie.

1.1.2 Naturalism and the Lockean Labour

I consider myself to be a philosophical naturalist, in the sense that I believe that if answers to the deepest questions of epistemology and metaphysics are to be found at all, they are to be found within Nature. And even if some of those questions can never be answered to the satisfaction of all enquirers, I believe that philosophers may usefully eliminate some of the wrong answers by adopting the methodology of the natural sciences.

Though I claim to be a naturalist, I adhere to the naturalist principle only by taking a broad view of what is meant by "Nature". I may therefore be disowned by those naturalists who equate Nature with the physical world, and naturalism with materialism or physicalism.

The philosophical naturalist position is strongly identified with Quine, who defends it in his famous essay "Epistemology Naturalized" (1969). Like many twentieth-century philosophers, Quine is also a physicalist, but he does not conflate the two positions. In a more recent paper, he expressly distinguishes the two:

> Naturalism is naturally associated with physicalism, or materialism. I do not equate them. …I do embrace physicalism as a scientific position, but I could be dissuaded of it on future scientific grounds without being dissuaded of naturalism. Quantum mechanics today, indeed, in its neoclassical or Copenhagen interpretation, has a distinctly mentalistic ring. (Quine 1995: 257)

Although Quine as a physicalist is unlikely to agree with the position I shall defend in this thesis, I think I fall within his concept of a naturalist.

Though I consider myself a naturalist, I am not a physicalist by any normal understanding of what is meant by "physical". As I shall argue in later chapters, I believe there is a sense in which human beings exercise free will that cannot be reconciled with a universe containing only matter and energy, or in which all physical events are necessitated or have their probabilities determined by prior physical events and states of affairs. Though that leads me to postulate entities or causes that are not physical, I would not say that such a view commits me to "supernatural" entities or causes. It merely requires a recognition that there is more to "Nature" than the particles, fields and forces currently identified by physics, or the entities postulated within the theories propounded by physics to explain that vast multiplicity of phenomena, from sub-atomic particles to galaxies, that do not involve the free will of conscious agents.

If the free will question is ever to be answered, I believe the answer will be provided by a future empirical science. But that is not to deny the important contribution that philosophy can make by clarifying the question that needs to be answered, and by eliminating those purported solutions that address the wrong question or fail to address properly the right question. Having come to philosophy late in life, my ambitions are modest. I shall be satisfied to emulate Locke's under-labourer, by " …*clearing Ground a little, and removing some of the Rubbish, that lies in the way to Knowledge"*. (Locke 1975: 10, italics in the original).

1.1.3 Divergent Aims of Philosophy

It seems to me that philosophy has two broadly divergent aims, which might be contrasted as descriptive and normative. While some philosophers may be able to address both of these aims, at least on different occasions, I suspect that most devote their careers largely to one or the other.

Wittgenstein is firmly in the descriptive camp. In the *Philosophical Investigations* , he observes:

> Philosophy may in no way interfere with the actual use of language; it can in the end only describe it. …
> It leaves everything as it is.(Wittgenstein 1953, para 124)
>
> Philosophy simply puts everything before us, and neither explains nor deduces anything. (*ibid.*, para 126).

Marx, as reported by Popper (1966, vol 2: .80) is in the opposite camp:

> The philosophers … have only interpreted the world in various ways; the point however is to change it.

Cartwright's philosophy is also motivated by normativity. In her introduction to *The Dappled World,* (1999: 5), she contrasts her own approach to the philosophy of science with that of Van Fraassen:

> Mine is the motive of a social engineer. ... Van Fraassen tells us that the foremost question in philosophy of science today is: how can the world be the way science says it is or represents it to be? I am interested in intervening, So I begin from a different question: how can the world be changed by science to make it the way it should be?

Though her aim is avowedly normative, to begin from the question of how to make the world "the way it should be" is to suppose that the way the world should be is antecedently known. In itself a specification of how the world should be is a normative enterprise, but if that specification is to be grounded in some kind of moral facts, the specification of those facts would be a descriptive task.

In the passages quoted, Wittgenstein *describes* what philosophy *does,* while Marx *prescribes* what it *ought* to do. Cartwright, on the other hand, legitimately *describes* her own aim as a *normative* aim. Although descriptive analysis can be employed to serve an ultimately normative end, a philosopher who asserts that the goal of philosophy *ought* to be purely descriptive may leave himself open to criticism.

1.1.4 Global Realism

I shall not fall into the trap of *prescribing* that philosophy *ought* to be descriptive, but I state here that my own preference is for descriptive philosophy. In stating that preference, I acknowledge that purely descriptive philosophy may be on shaky ground, for it presupposes there is some reality to be described. Nevertheless, in this thesis, I adopt a perspective of global realism, and assume not only that there is a reality to be described, but further that there is a single reality, about whose overall constitution and nature it is at least meaningful to speculate. Against the sceptic, I doubt that any useful version of realism can be proven correct, but by assuming that some kind of mind-independent reality underlies the experiences people have of the world, it is at least possible to postulate categories, models and regularities that sometimes succeed in forming a basis for predicting future experiences.

The position I call "global realism" is criticised by Cartwright (1999), who calls it "fundamentalism". Cartwright is willing to endorse a local form of realism on pragmatic grounds. It is only by admitting, at least defeasibly, the truth of empirically verified local regularities or laws that we can form concepts and learn to interact with the world. She speaks of a "patchwork" of local laws, each valid within its own domain and perhaps also valid in a few other domains in which it has been established to work. But Cartwright refuses to extrapolate or extend the laws derived in one domain to other domains, or to assume that the laws we find and corroborate within any domain or set of domains have universal application. Those motivated towards such extrapolation or universalisation are the "fundamentalists" she has in her sights.

By Cartwright's account, I am a "fundamentalist", although I think that term is unfortunately pejorative. To me, fundamentalists are people who not only believe that there is some underlying truth about a subject, but who also believe that they possess that truth, and therefore refuse to consider any contrary views. In the early twenty-first century, the label of fundamentalist has come to be associated with people who not only refuse to consider views contrary to their own, but are motivated to kill those who disagree with them and anyone else who gets in the way. I am not a fundamentalist in that sense.

Global realism is important to my thesis, as I aim to draw conclusions about the nature of the world, and in particular about how free will in the sense to be made precise in Chapter 2 may be reconciled with the features of reality that are described by physics. For that project, I need to offer some hypotheses about the world, and some of those hypotheses will have a broad domain.

Not only do I need to adopt the perspective of global realism to support my thesis, but I find that position more plausible than Cartwright's local realism. If the laws of nature are localised as Cartwright suggests, then each of those laws would have to have a domain of applicability. Unless we could say whether a particular system is subject to a given local law, we would not have a law at all. There would be no regularities. That means that whether or not we could know them, there would have to be metalaws defining the domain of each local law. If so, then a conjunction of all the local laws together with all the metalaws defining their domains would constitute a theory of the wider world. Such a theory would be far more complex than a theory based on universal laws, and to me less attractive for that reason.

Though I suspect that one can never prove the correctness of global realism, neither can that assumption be disproved. But if one is prepared to accept the premises that lead Cartwright to local realism, then I think that inductive evidence in support of global realism can be found. The history of science contains many examples of theories developed in a narrow domain that have later been found to be applicable in other domains, if not universally. That topic is explored in Chapter 6 of this thesis.

Cartwright concedes that neither her dappled world nor the "fundamentalist's" world of universal laws is conclusively proven to be the correct model, though she does claim that her dappled world view is the one "best supported by the evidence" (1999: 12). Her principal reason for choosing the dappled world view rather than simply reserving judgement arises from her normative approach to philosophy, as mentioned in Section 1.1.3:

> Why then choose at all? Or, why not choose the risky option, the world of unity, simplicity and universality? If nothing further were at stake, I should not be particularly concerned about whether we believe in a ruly world or an unruly one, for, not prizing the purity of our affirmations, I am not afraid that we might hold false beliefs. The problem is that our beliefs about the structure of the world go hand-in-

> hand with the methodologies we adopt to study it. The worry is not so much that we will adopt wrong images with which to represent the world, but rather that we will choose wrong tools with which to change it. (Cartwright 1999: 12)

I willingly allow that Cartwright's choice best serves her goal of changing the world to the way she would like it to be. Conversely, my own choice of global realism better serves my Lockean goal of clearing the way to a better knowledge of the way the world is. No more than Cartwright am I *afraid* that I might hold false beliefs. I am sure many of my beliefs *are* false. But I think people generally prefer true beliefs to false beliefs, and on some matters it is important to us that we acquire and maintain as true a set of beliefs as possible. A desire to have a better knowledge about the way the world is drives my descriptive approach to philosophy.

1.2 Various "Free-Will Problems"

The approach I take in this thesis is, therefore, a descriptive one. I aim to explore and elucidate how it can be that I have the inescapable impression that on numerous occasions there are genuine alternatives available to me, and that I seem to be able to make a free choice as to which of those alternatives I adopt. Not only do *I* have such an impression, but I think it is safe to say that just about all philosophers writing about "free will" agree that people in general have such impressions, though some say such impressions are illusory. Further, the way in which people interact with each other in everyday life seems to take it for granted that each of us has, in our dealings with each other and the world, an ability to make choices as to how we shall act, and what we shall believe.

Although a multitude of papers have been written about "the free-will problem", no single problem answers to that description. At the risk of begging the question, it seems that we are able to choose our own "free-will problem", depending on what assumptions we wish to make.

For a Cartesian, the free-will problem is to explain how an autonomous immaterial "self" is able to interact with the material realm and to affect the material world. Descartes speculated that mind-body interaction took place in the pineal gland (1988, Articles 31-34), but his speculations have been discounted by empirical science. Substance dualism is no longer a popular position, though Eccles and his collaborators continue to postulate a Cartesian self, and to offer hypotheses as to how that self controls the physical brain (Popper and Eccles 1977, Eccles 1994, Beck and Eccles 2003). Such speculations fall partly within empirical science, but a philosophical question remains as to the nature or constitution of the self, if it is not a material entity. The work of Eccles and his collaborators is discussed in Section 6.3.

For many contemporary metaphysicians, the problem of free will is to account for the appearance of free will in a physicalist world. If physics provides the model for a complete description of reality as many believe it does, it is not apparent how agents could make free choices. Classical physics is deterministic. Given a complete description of the world at one instant, the laws of physics are supposed to determine its state at any future (or past) instant. There is no way that an agent could affect future states of the world unless that agent's behaviour were itself part of the structure determined by the physical past, and that seems inconsistent with the naïve view that some of our actions are "up to us".

Unlike classical physics, quantum mechanics gives an indeterministic description of the physical domain, but the indeterminism of quantum mechanics does not solve the physicalist's free will problem. Where quantum mechanics predicts a range of two or more possible outcomes, which one of those outcomes actually occurs is purely random in a physical system. And a random event would seem no more "up to us" than one determined by external factors.

In Chapter 5, I shall argue that a quantum-theoretical description of the physical world is essential to an account of free will, but the mere randomness allowed by quantum mechanics does not assist the physicalist. Philosophers determined[1] to maintain their belief in physicalism must give some account of the undeniable fact that people *seem* to have free will, and mostly act as if they believe they have it. Physicalists may respond either by denying that people really have free

[1] Ambiguity intended.

will, or by offering some account of "free will" that is compatible with their physicalist view of reality.

Yet another free will problem is of concern to moral philosophers. It is generally agreed that if people are to be held accountable for even some of their actions, there must be some sense in which society can treat such actions as freely chosen. Not only in abstract philosophical discussions but in judicial reasoning, in political debate and in everyday social interaction, people are apt to hold others accountable for their actions according to whether they "could have done otherwise" than they did. In Section 2.4 I shall argue that a person's acting freely in the moral sense presupposes that that person has a free choice in the metaphysical sense that the action is neither random nor determined by external circumstances, but many philosophers strive to give an account of acting freely that justifies the attribution of praise and blame, even if the agent's behaviour is caused by his upbringing, character or dispositions. Important as these moral issues are, they are not the issues I address in the main chapters of this thesis. In Section 1.3, I shall argue that the metaphysical question of free will is better analysed in isolation from the moral question.

1.2.1 The Metaphysical Concept of Free Will

In chapter 2, I shall specify a metaphysical concept of free will, and I shall do so in a context divorced from any considerations of right or wrong, praise or blame. It *seems* to me that there are many occasions on which I have a free choice in what I do. Implicitly, it seems that some events in my future are neither completely determined by circumstances beyond my control, nor attributable to pure chance which is equally beyond my control.

One partial explanation for my *seeming* to have choices is that the future is not fully determined, and that I do in fact have choices. That is a proposition which I shall defend, but it is by no means essential to an explanation for the way things seem. A well-worn example is that the sun *seems* to travel across the sky, yet most people now accept that the best explanation for this phenomenon is the Earth's rotation on its axis.

That the sun *seems* to move across the sky is a belief shared by most people, including those with sufficient education to know that in fact the Earth rotates. In contrast, few, if any individuals actually "imagine that they have an earthenware head or are nothing but pumpkins or are made of glass", to borrow Descartes' (1911) example. If my belief that I seem to have choices were mine alone, or shared with only a few other aberrant individuals, the source of my belief might be something peculiar to myself, and of no general philosophical interest. I claim, however, that a belief that they at least seem to have choices about their future is shared by most of humanity. Support for the universality of such a belief may be found in human languages. In Chapter 2, I shall refer to semantic data showing how that belief is deeply embedded in the English language.

Not only do I claim that people *believe* they have a capacity to make choices. I also argue that such a capacity is exhibited and exercised by normal human agents in their everyday life. For me, therefore, the problem of free will is to accommodate the capacity of human agents to exercise choices that affect the physical domain within a globally real conception of Nature. I shall defend the claim that such accommodation requires a broader conception of Nature than is offered by physics alone.

1.3 Free Will, Praise and Blame

As mentioned in Section 1.2, many philosophers approach the questions of whether and in what sense we have free will as dealing with ethical issues. It is widely though not universally taken for granted that persons should not be held to account for events or states of affairs over which they had no control, and that seems to require that in some sense or other the recipient of praise or blame must have acted freely. Those committed either to universal determinism or to determinism relieved only by instances of pure chance must therefore find a plausible notion of free action which is compatible with their causal commitments.

Although I agree that persons should not be held accountable for events and states of affairs beyond their control, I also contend that most of the actions that are freely performed in every day life have no moral consequences at all. There may be a

moral question as to whether I should enjoy a banquet while someone else is starving, but if I am already seated at the table and constrained by courtesy towards my hosts to eat what is on my plate, it is hard to see any moral issue in whether I should eat my peas before my carrots or my carrots before my peas. Yet it seems to me that I have just as free a choice in those circumstances as in reaching some weighty decision, laden with moral consequences.

In the chapters to follow, I shall attempt to keep the metaphysical question of free will separate from the moral question for two reasons. In the first place, solving the moral question would not account for free will in those trivial cases in which no moral consequences are discernible. In the second place, people tend to have strong views about moral questions. Reaching the "right" conclusion about a moral question "matters" to people in ways other than a simple desire to understand how the world works. Just as Cartwright's philosophy is guided by her desires about the way the world should be, discussions of free will with moral consequences can be influenced by the interlocutors' approval or disapproval of an agent's action. Even if such considerations do not lead to a wrong answer, they risk obscuring the right one.

In a book entitled *Problems of Ethics*, Moritz Schlick approaches the free will problem with expressions of reluctance. Though his conclusions on free will are very different from mine, he too urges the separation of moral issues from metaphysics:

> Desire for the truth is the only appropriate inspiration for the thinker when he philosophizes; otherwise his thoughts run the danger of being led astray by his feelings. His wishes, hopes, and fears threaten to encroach upon that objectivity which is the necessary presupposition of all honest enquiry. Of course the prophet and the investigator can be one and the same person; but … whoever mixes the two problems will solve neither. (Schlick 1962: 2).

Schlick is a determinist, whose solution to what he calls the "pseudo-problem" of free will is discussed in Section 3.5.7. As will be explained there, the free will he reconciles with determinism is not the metaphysical free will that concerns me. Schlick expressly denies the feeling of choice that motivates my metaphysical concept of free will. By his intuition, the feeling that one could also have acted otherwise "… never says that I could also have willed something else". I do not dispute Schlick's report of his own intuition, but can only say that I do not share that intuition. Unless

there are two kinds of people in the world, respectively with and without free will in the metaphysical sense, one of us must be wrong about the "feeling of choice". The truth of the matter cannot be settled merely by an appeal to intuition, but I shall offer two arguments in defence of my own position in Section 2.3.

1.3.1 The Campbell – Smart Exchange

The mid-twentieth century sees an exchange in *Mind* between C.A. Campbell (1951, 1963) and J.J.C. Smart (1961), approaching the concept of free will as a basis for moral responsibility. Though both authors are motivated by moral considerations, they draw metaphysical conclusions.

Campbell's earlier paper is primarily a reply to logical positivists like Schlick, and much of it responds to Schlick's equation of responsibility with aptness for punishment, to be discussed in Section 3.5.7. But Campbell also introduces a concept similar to the sense of metaphysical free will I shall define in Section 2.2. He claims that at least some human actions are the result of a genuine choice, in the sense that two or more outcomes exist for the agent, none of which is causally determined. Campbell introduces the term "contra-causal freedom" to refer to that sense of freedom, and he claims that contra-causal freedom is a necessary condition for attribution of moral responsibility. At least, he contends, that is the view of "the common man", and he goes on to argue that no hypothetical understanding of "could have done otherwise" captures the sense of "could" required for attribution of responsibility. I agree that human agents possess what Campbell calls "contra-causal" freedom, and as discussed in Section 2.4, I agree that its possession is a necessary condition for moral responsibility. Like Campbell, I shall also reject hypothetical understandings of "could have done otherwise" in Section 3.5.2. I find the term "contra-causal freedom" infelicitous, however, because I think exercises of what I prefer to call "metaphysical free will" are better seen as instances of causation, though not the ordinary causation by prior events according to deterministic or probabilistic laws.

Campbell speculates on what leads other philosophers to reject the common man's understanding of moral responsibility. In addition to the positivists' need to confine discussion to statements that are in principle subject to verifiability, he concludes that his opponents are in the grip of a "predisposing influence", "extrinsic to their specific arguments", namely "the conviction that there just *is* no contra-causal freedom" (Campbell 1951: 459). He attributes that conviction in part to a general presumption in favour of universal causation, but in particular to "… the common assumption of social intercourse that our acquaintances will act 'in character'".

Campbell does not here take issue with the "general presumption in favour of universal causation", but he goes on to deal with the argument about acting 'in character'. Campbell concedes that free will "does not operate in practical situations in which no conflict arises … between …'duty' and …'strongest desire'. In such situations, he says the agent's conduct is predictable from his character, but it is in the other situations, where 'duty' and 'strongest desire' conflict that free will manifests itself. Here, I disagree with Campbell. As stated in Section 1.2.1, I believe there are many exercises of free will that have no moral dimensions whatever, but the cases in which we judge the behaviour of others are those that do have a moral dimension. I would say that free will could equally be said to operate in cases where 'duty' and 'strongest desire' coincide. In those cases, the agent still has the ability to act perversely, but has no motivation to deviate from the choice that both duty and desire indicate. The agent freely chooses, and his choice is predictable to anyone knowing his character and obligations.

Where 'duty' and 'strongest desire' coincide, the agent's choice is likely to be predictable, but still only at some levels of description. A mother's 'duty' and 'strongest desire' both influence her to feed her child a healthy diet, but that does not mean it is not within her power to let the child go hungry or feed it junk food on some occasion. And even someone intimately aware of the mother's character and upbringing, predicting with near certainty that she will feed the child on some occasion cannot predict whether she will give the child and apple or a pear, or whether she will tender the morsel with her right or her left hand, or whether she will serve it in the Bunnykins® or the Wiggles® dish.

Cases where 'duty' and 'strongest desire' seem to conflict seem to be different because others judging the agent's choice are not in a position to weigh the competing

influences. I believe that for most moral agents, a desire to "do what is right" or to receive the approval of our fellows is an important member of the highly complex set of desires which make up our characters. If our actions were determined as the resultant of a set of competing desires, then "duty" and "strongest desire" may oppose each other, but the action would be a resultant of all relevant desires, strong and weak, conscious and unconscious. But to say our actions are determined as the resultant of a set of competing desires is to deny that in some sense the action is "up to us". When making a morally important decision, what we *seem* to do is to deliberate, sometimes agonisingly and at great length, attempting to weigh our desires and duties against each other. At the end of that deliberation, we reach a decision, and we act in accordance with that decision. But the decision in such cases does seem to be the agent's decision, rather than the result of some deterministic, algorithmic process by which competing desires and duties are objectively balanced against each other. I believe Campbell concedes too much when he does not challenge the concept of universal causation.

Smart's (1961) response to Campbell illustrates how the mixing of metaphysical and moral issues can arouse the passions. Smart sets out "…to refute the so-called 'libertarian' theory of free-will". His attack is driven by a repugnance for the social consequences he sees in the acceptance of the theory, and his paper abounds with emotive language[2]. He likens Campbell's position to that of "the notorious Marquis de Sade". Such rhetorical devices can have no bearing on the metaphysical question of free will, though Smart's conclusion is a metaphysical one.

Stripped of its rhetoric, Smart's paper claims to show that contra-causal freedom is logically impossible. He does this by defining "unbroken causal continuity" and "pure chance" each in terms of a superhuman (Laplacian) calculator's ability to predict events, given full knowledge of the background conditions. On his definitions, the two concepts are contradictories, admitting no third possibility. In my view, the argument is unsound, since Smart's definition of "pure chance" includes just

---

[2] For example:

> When, in nineteenth-century England, the rich man brushed aside all consideration for his unsuccessful rivals in the battle for wealth and position, and looking at them as they starved in the gutter said to himself, "Well, they had the same opportunities as I had. If I took more advantage of them than they did, that is not my fault but theirs", he was most probably not only callous but (as I shall try to show) metaphysically confused. (Smart 1961: 291-2)

those circumstances in which his libertarian opponent would claim the agent has contra-causal freedom. If a deliberating agent has not yet made up his mind, neither he nor the Laplacian calculator can predict the decision, but the libertarian would deny that the decision, when made, is attributable to "pure chance". Thus, I believe Smart begs the question.

Smart agrees that freedom exists, but in a sense which is nonetheless compatible with determinism, and similar to that of Schlick. No human actions are free from causal determinism, but we are nonetheless justified to some extent in attributing praise or blame, by the assumed desirability of encouraging or discouraging behaviour conforming to some concept of morality.

In a brief rejoinder, Campbell also argues that Smart begs the question against him, and suggests an amendment to Smart's definition of "pure chance", expressly to exclude cases of contra-causal freedom. He denies that would equally beg the question:

> Our revised [definition] by no means entails that there actually are instances of "contra-causal freedom". It merely takes account of the possibility that there *may* be. Smart's [definition] assumes that there *can't* be. (Campbell 1963: 410)

The rest of Campbell's rejoinder addresses specific points made by Smart about details of Campbell's original paper, and does not seem to take the debate much further. I think Campbell and Smart differ over an important metaphysical question, but that question could have been examined more clearly if the separate moral question had been left out of the debate.

1.3.2 Strawson's Freedom and Resentment

P.F. Strawson is another philosopher whose interest in free will is motivated by moral concerns. In a famous paper entitled "Freedom and Resentment" (1962), he sets out to identify a concept of freedom, sufficient to justify society's practices of reward and punishment, but not incompatible with what people mean when they talk about determinism. Previous compatibilist accounts had defined freedom in negative terms

not conflicting with determinism, but for Strawson, such accounts miss an important aspect of the concept of freedom:

> [W]hat 'freedom' means [according to such accounts] is nothing but the absence of certain conditions the presence of which would make moral condemnation or punishment inappropriate. They have in mind conditions like compulsion by another, or innate incapacity, or insanity, or other less extreme forms of psychological disorder, or the existence of circumstances in which the making of any other choice would be morally inadmissible or would be too much to expect of any man. (P.F. Strawson 1962: 60)

Strawson says that the utility of social practices of praise and blame is not impaired by something like determinism, but he argues that human intuitions require something more, namely that the attributee of praise or blame "must really deserve it", and that "in the case at least where he is blamed for a positive act rather than an omission, [there need be] a genuinely free identification of the will with the act." (*ibid*.: 61).

Strawson coins the terms "optimist" and "pessimist" respectively for those groups of philosophers who believe that determinism can be reconciled with a moral intuition-satisfying notion of freedom, and those who do not. Although he denies allegiance to either camp on the ground that he does not really know what determinism is, he claims that the intuitions about desert and authorship are explained in terms of what he calls "our ordinary reactive attitudes" towards ourselves and others viewed as agents.

According to Strawson, our attributions of praise and blame are not merely tools society uses to control its members, but expressions of deeply-rooted attitudes. Those attitudes could not be renounced even if we were prepared to accept the theoretical truth of determinism, and so the truth or otherwise of determinism is irrelevant to question of what constitutes freedom for moral purposes.

The "pessimist's" intuition for "desert" can be accommodated in a deterministic world, but only if he will "surrender … his metaphysics" (*ibid*.: 78). Strawson manages to avoid what he calls "the obscure and panicky metaphysics of libertarianism" (*ibid*.: 80), but only by refusing to address the metaphysical problem. Unlike Strawson, my primary interest is in the metaphysical problem. I suggest it is possible to address that problem without panicking, and by doing so we may reduce

the obscurity, and perhaps gain useful insights into the causal fabric of the world. As a by-product, we may also come closer to solving the problem that interests Strawson. The possibility that the solution may turn out to be 'libertarian' is no reason to refuse to consider it.

1.4 Degrees of Freedom

In Chapter 2 I shall present and defend arguments that human agents have free will in the sense that some of our acts are "up to us". Though it is not essential to my existential claim in that chapter, nor to the conclusions that follow from it in later chapters, I think it can be argued further that in all those acts that human beings perform consciously, there is at least a residual element of free choice.

I shall say more about what I understand by "act" in Chapter 2. For the present discussion, I rely on the intuitive concept of acts being physical events or linked sets of events involving some movement or change in a person's body, for which normal speakers of English would regard the person as somehow being causally responsible. I think by that standard, raising an arm and whistling a tune count as acts, whereas dying, perspiring and (involuntarily) sneezing do not. These examples suggest that common usage imputes a sense of voluntariness into the concept of an act. That is an empirical claim for which I offer no evidence, though such evidence could no doubt be collected by proper questioning of a sufficient sample of English speakers. As the contention is not essential to the main arguments in my thesis, I ask the reader's indulgence.

If the concept of an act requires some element of voluntariness, what can be said about acting under duress? Among those cases in which social custom and criminal laws excuse acts that would otherwise be condemned are those where it is accepted that the agent did not act freely in the socially or legally relevant sense: where we would agree that the agent "could not have done otherwise".

The phrase "could have done otherwise" occurs a lot in discussions of free will. With a few conspicuous exceptions[3], most philosophers would agree that unless an agent "could have done otherwise" in *some* sense of those words, he or she did not act freely. I shall return to that phrase in Section 2.2.5, after a semantic analysis of various uses of the modal verbs CAN and COULD[4]. My claim in the present section is that in any event properly described as the act of some person, there is some element of voluntariness, identifiable with the recognition that in *some* sense, that person could have done otherwise.

Consider the hackneyed example of a bank teller forced at gunpoint to open a safe. Presented with the facts of the case, a jury could reasonably find that the teller could not have done otherwise. For the purposes of legal responsibility, it would be perverse to say that the teller had a choice, and that he could have chosen to be shot. Yet for other purposes, it would be true that the teller had a choice. However unpalatable the alternative, and however unreasonable it would be to expect him to have done so, the teller could have refused, or attempted to disarm the robber, or tried to run away.

More generally, and however strictly some agent might be constrained by external circumstances to perform some act, I claim that such an agent retains a degree of freedom, if not on whether to perform that act, at least in how the act is performed. Our teller under duress opens the safe with his right hand, but he could have opened it with his left. If forced to use his right hand, he might have chosen which fingers to use, or how much pressure to exert on the knob, or how rapidly to turn the knob. I contend that for any genuine act performed, it could truthfully be said the agent could have done "otherwise", if the token act as performed is given a

---

[3] Dennett (1984, p131ff., 1984a) is one of the exceptions, but he describes the "could have done otherwise principle" as one of "a few calm islands of near unanimity" in "the midst of all the discord and disagreement among philosophers about free will". Dennett quotes van Inwagen:

> Almost all philosophers agree that a necessary condition for holding an agent responsible for an act is believing that the agent *could have* refrained from performing that act. (van Inwagen 1975, p. 189)

[4] Frankfurt (1969) offers some ingenious thought experiments in which the concept "could have done otherwise" comes apart from the notion of choice that I think characterises free will in the metaphysical sense. At worst, I think Frankfurt's examples show that a person whose neural processes are intercepted by some mischievous or malevolent demon does not have free will. That does not impair my metaphysical claim that normal agents exhibit free will on at least some occasions.

sufficiently "fragile" description, according to Lewis's (2000) terminology. That is not to suggest that agents reflect so deeply on each token act they perform, but merely that they normally have a capacity to do so. Common experience suggests that in any project we are undertaking, if we allow ourselves to reflect, we can find some trivial way in which to stamp our authorship on the deed. Even the guard standing strictly to attention at Buckingham Palace can wiggle his toes inside his boots or perform mental arithmetic to reassure himself he is still alive. If there were an extreme case of duress where the "agent" was so much in fear that he was unable to control any aspect of the event he was compelled to bring about, I would be prepared to say that event was not his act at all, and that he was not an agent but an instrument of the coercer.

1.5 The Indispensability and Limitations of Language

In attempting a philosophical explanation, we cannot escape language. Language in some form or another — I include sign language and mathematical language — is indispensable for *communicating* abstract or complex ideas, and I suspect that at least for beings with minds like humans, it is also indispensable for *reasoning about* abstract ideas.

I do not claim that language is necessary for beings to exercise a degree of free choice, and to influence their environment. I would not rule out the possibility that a dog can sometimes choose (say) to which of two people he will bring a retrieved ball, although I doubt that he spends much time reflecting about that choice. Because we cannot communicate with dogs so effectively as with each other — especially about abstract things — it will be harder to defend a thesis about free choice among non-human beings, and I shall not attempt to do so. I shall, however, draw heavily on the resources of language to draw inferences about how people at large view their capacity as free agents, to distinguish among several different conceptions of free will, and to draw inferences about the causal structure of the natural world.

Though the deep structure and semantic categories of language may provide evidence of some widely-shared conceptions of reality, I doubt that language could ever prove the truth of those conceptions. People may be systematically mistaken

about the nature that underlies their apparent experiences. But I believe the best evidence any of us can have for the appearance of free will is our own experience as agents in the world. Language allows me to supplement my own direct experience with evidence of the experience of others, either as directly reported by them, or as embedded over millennia in the semantic categories of volition that underlie a major part of most if not all human languages.

1.6 Outline of Succeeding Chapters

In Chapter 2, I attempt to make precise the concept of free will that I plan to discuss in later chapters. Taking the English language as one means by which many humans communicate their experiences of the world, I show that that language has developed a complex system of modal verbs to express a variety of senses of ability and possibility that its speakers find important. Referring to linguistic research, I develop the resources to distinguish a particular sense of the verb CAN and its past tense form *could* that may be used to describe what I call "metaphysical free will", and I distinguish that sense of the modal verb from other senses of the forms *can* and *could* that figure in descriptions of other senses of freedom.

I argue that metaphysical free will in the sense described is not just a curiosity, but an important capacity exhibited by human agents. Although exercises of metaphysical free will *per se* have no necessary moral consequences, I argue that the possession of that capacity is presupposed in the other senses of freedom that are important in human affairs, including the freedom required for moral responsibility.

Finally in Chapter 2, I offer an independent argument from evolutionary theory supporting the claim that human agents have metaphysical free will, and showing that our belief that we have such a capacity cannot be dismissed as an illusion.

In Chapter 3, I argue against the widely held view that "free will" in every important sense of the phrase is compatible with determinism. I begin by examining the well-known "consequence argument" espoused by van Inwagen and others, that purports to demonstrate that a metaphysical conception of free will according choice to agents is incompatible with determinism. I also consider some criticisms of that

argument, and some refinements of the original consequence argument that have been offered in response to those criticisms.

Although I find the consequence argument persuasive as it stands, I believe much of the debate it inspires has been led astray by equivocation over the kind of modality expressed in its premises. I therefore offer an alternative argument to show that metaphysical free will in the form to be made precise in Chapter 2 is not only incompatible with determinism, but more importantly, incompatible with causal closure of the physical domain. The distinction is important because although most compatibilists would concede that the world as revealed by quantum mechanics is not deterministic, they assume that if free will is compatible with determinism, it is *a fortiori* compatible with the kind of indeterminism that quantum mechanics admits. But incompatibility with determinism does not entail incompatibility with causal closure of the physical domain. An incompatibilist therefore has the harder task of showing that free will in an important sense is incompatible not only with an artificial deterministic world, but with the real world that current physics tells us we inhabit.

Because compatibilism is probably the dominant view among philosophers, I continue in Chapter 3 by considering some well-known arguments for compatibilism of "free will" with determinism or causal closure of the physical domain. By analysing the modal language used, I show that the concepts of free will to which those arguments refer are distinguishable from the concept of metaphysical free will specified in Chapter 2.

If it be accepted that human agents have the capacity of metaphysical free will specified in Chapter 2, and if that capacity is incompatible with causal closure of the physical domain, a global realist is forced to abandon a belief in causal closure. Belief in the causal closure of the physical domain seems to be deeply entrenched among philosophers, and in Chapter 4, I search for the source of that belief. I surmise that many philosophers either take the belief on authority, or are persuaded by an inductive argument based on the historic success of physics to explain various phenomena that had once been believed to lie outside its causal domain.

In Chapter 4, I also examine and reject some arguments to the effect that failure of physical causal closure would contradict well-established physical laws, including the conservation laws and the second law of thermodynamics. I conclude there is no compelling reason to insist on causal closure in face of that doctrine's

incompatibility with the capacity for metaphysical free will, the reality of which I claim is better supported than the causal closure principle.

Chapter 5 begins with a discussion of the boundary between physics and metaphysics. Although the free will question presently falls within the domain of metaphysics, I suggest that philosophers can be well guided by the approaches taken by empirical science.

I then compare the free will problem with one of the enduring mysteries of physics: namely the measurement problem of quantum mechanics. Though the mere intractability of these two problems does not imply that they are related, I give reasons for thinking that the account of physical reality given by quantum theory is the one most likely to be reconcilable with metaphysical free will. I conclude the chapter by discussing the apparent reliance of some interpretations of quantum mechanics on a conscious mind, external to the physical systems quantum mechanics purports to describe.

In Chapter 6, I discuss some recent and continuing research by philosophers and empirical scientists who are working to explain the phenomenon of consciousness and particularly the capacity for free will, within the indeterminism inherent in quantum mechanics. Though consciousness is conceivable without free will, I cannot imagine free will without consciousness, and any advances in explaining consciousness could advance the search for an explanation of free will. Though not endorsing any one of these research projects, I take encouragement from the existence of several avenues of investigation and development. I compare and contrast various projects, and offer a speculative and tentative solution that combines features from two of them.

Lastly, I look at the history of physics, and its partial success in combining initially diverse phenomena into unified theories of ever-greater generality. I also refer to the superstring and M-theories proposed by some physicists. These theories are presently more metaphysical than physical, but are intended to provide a unified account of all physical phenomena in terms of “strings” or multi-dimensional “branes”. Even if superstring or M-theories in future receive sufficient empirical support to gain acceptance as theories of physics, they do not currently offer any account of either consciousness or the capacity for free will. I nonetheless suggest that the methodological approach taken towards unifying physical theories might

usefully be employed to find an overarching theory of mind and physics. The ultimate theory of free will is best sought within an extended conception of Nature, of which the physical domain is but one proper part.

CHAPTER 2

METAPHYSICAL FREE WILL

2.1: The Semantics of CAN and COULD

2.1.1 Introduction

The words "can" and "could" permeate the philosophical literature of "free will". Many philosophers have addressed the "free will problem" (Section 1.2) by analysing expressions such as "could have done otherwise". The argument I shall develop for the incompatibility of free action with causal closure of the physical domain attributes to agents an ability to perform actions, most naturally expressed in English through the auxiliary verb CAN[5]. In the first part of this chapter, I shall examine the grammar and semantics of the (grammatically) modal verb CAN and its preterite[6] form *could*, as used in contexts implying or assuming free agency. In doing so, I shall distinguish those uses from other uses of CAN and its cognates.

Borrowing from linguistic taxonomies, I shall identify several distinct semantic categories realised by various uses of *can* and *could*. By reference to those semantic categories, I shall describe a particular metaphysical sense of "free will" that

---

[5] In this Chapter, I follow the linguists' convention of using small capitals for lexical items such as BE, and italics to refer to individual forms of those items (such as *am, is, are, was, were, be, being, been*). Modal verbs have a restricted number of forms, and the only forms of CAN seem to be *can*, *could*, and their negations *can't, cannot* and *couldn't*.

[6] I choose the somewhat archaic name "preterite" rather than "past" or "past tense" to emphasise that this form of the verb can be used for purposes other than denoting past time.

I claim is exhibited by free agents. I shall show how this metaphysical free will differs from some other important senses of free will that have been analysed by philosophers, but I shall argue that attributing freedom in those other senses to an agent tacitly presumes that the agent has freedom in the metaphysical sense.

Not only is metaphysical free will presupposed in day-to-day human activities, but I shall argue that it is a real capacity of human agents, and cannot be explained away as an illusion. I shall offer arguments based on the structure of language and on natural selection, and I shall respond to Wegner's (2002) claim that the experience of conscious will, though essential to human society, is nonetheless illusory.

2.1.2 The English Modal Verbs

Grammatically, CAN is one of the English modal verbs, which together form a subclass of the closed class of auxiliary verbs. In a comprehensive linguistic study of the English modal verbs, Palmer notes that they are distinguished from other auxiliaries, and from verbs in general, by the following syntactical features:

-- No *-s* form of the third person singular. (No **mays, *cans* etc.)
-- No non-finite forms (infinitives, past and present participles).
-- No co-occurrence. (No **He may will come* etc.) (Palmer 1990: 4)

CAN, like WILL, MAY, SHALL, COULD and MUST is a paradigmatically modal verb. Although the English modal verbs are easily identified either by their syntactic peculiarities or simply by enumeration, their semantic function or functions are less easily categorised.

2.1.3 The Epistemic/Root Distinction

One major distinction recognised by most linguists is between the use of modal verbs to express epistemic modality, and their use to express non-epistemic or "root" modality. Epistemic modality signals a speaker's or author's attitude to the

proposition expressed by the verbal complement: typically the extent to which the speaker judges that proposition to be true. "Root" modality expresses features internal to the proposition, such as the necessity or possibility of the state of affairs it describes.

Although clear examples can be given of epistemic and root modality, such examples merely represent the endpoints of a cline. Most of the modal verbs can be used to express either epistemic or root modality, or some degree of each. Syntactically identical sentences can perform either function, and the speaker's meaning must be judged from context or prosody. For example:

John *could* be in his office.

is ambiguous. It could mean "For all I know, he is there", or it could mean there are no circumstances preventing his being there. The context in which the utterance is made would usually resolve the ambiguity. Among competent speakers of English, the ambiguity is often resolved by the way the sentence is spoken. In both the epistemic and the root uses, primary emphasis would be placed on the word *could*, but as the reader is invited to verify, a clearly epistemic utterance is likely to have rising intonation at the end of the sentence, whereas a clearly root utterance is likely to have a falling intonation:

For all I know, he *could* be in his office.

vs.

He got back to Sydney yesterday, so he *could* be in his office.

In the first example, the speaker expresses his doubt as to whether or not John is in the office. In the second, he confidently asserts the proposition that John's being in his office is possible.

Although the epistemic/root distinction is a useful one, many uses of modal verbs contain a degree of each. Thus, although the first of the pair just above is primarily epistemic, it is implicit in the speaker's belief that the speaker is unaware of any circumstance that would prevent John from being in his office, such as his being in the garden. And although the second is primarily root, it relies on the speaker's belief that John is back in Sydney.

2.1.4 Aspect in Verbal Groups

Different languages use various systems to place or relate events and states of affairs within time as perceived by speakers and hearers. English has a complex system of tenses, primarily to reflect the temporal relation between the event reported and the time of reporting, but also capable of expressing temporal relations amongst events, and of reflecting whether events and states of affairs are continuing, completed, or of a recurrent kind.

A feature well developed in English is the contrast between its simple present tense, as in "I walk" and the "progressive or continuous" present tense as in "I am walking". Another feature exhibited in many European languages is the contrast between the simple past (preterite) tense as in "I walked" and the past perfect tense as in "I have walked". Halliday and Matthiessen (2004: 337-354) account for such contrasts within a recursive tense system, but many other linguists invoke a separate system of aspect. According to Crystal (1995: 225):

> Aspect refers to how the time of the action of the verb is regarded – such as whether it is complete, in progress or showing duration.

Crystal recognises two types of aspectual contrast, the first between perfective and imperfective (I walked vs. I have walked) and the second between simple and progressive (I walk/walked vs. I am/was walking).

Either within its present tense system, or in a separate aspect system, English has a syntactic contrast between the progressive form "I am walking" and the simple form "I walk". The progressive form refers to an event that is taking place, and is not completed, at the time of utterance. The "simple" present form of a verb is used for events treated as instantaneous, or at least in circumstances where the duration of the event is not significant, as in a football commentary: "Socrates picks up the ball and passes it to Plato, who scores a goal". The simple present form of an event verb is also used when referring not to a single occurrence of an event token, but to a repeated or habitual occurrence of an event type, as in "I walk to work on Tuesdays".

Conceptually, a state of affairs cannot be instantaneous; it must have duration. Yet the simple and progressive forms can each be used to describe states of affairs, and competent speakers of English tacitly recognise a difference in meaning. We are more likely to say that the Cenotaph "stands in Martin Place" than to say that it "is standing in Martin Place", but less likely to say John "stands on the Post Office steps" than to say he "is standing" there. The semantic distinction here is aspectual. The Cenotaph's standing in Martin Place is a state of affairs presumed to have indefinite duration, whereas John's standing there is presumed to have a finite duration. If we are told "John stands on the Post Office Steps", we are likely to infer a habitual context, as in "John stands on the Post Office Steps each Anzac Day". The progressive form seems appropriate for current states and events whose finite duration is signified, while the simple form is used for both for events of insignificant duration, and for habitual events and current states whose duration, if not eternal, at least is indefinitely long. Such observations prompt Leech (1987: 19) to observe that "the Progressive *stretches* the time span of an 'event verb' but *compresses* the time span of a 'state verb'".

2.1.5 Semantic Aspect

Although the English present tense has a dichotomy of syntactic forms, those forms do not map neatly onto two semantic categories. To reflect better the temporal semantics of events and states of affairs, Coates (1983: 99) suggests a useful trichotomy of aspects. Verbs used to refer to discrete events are said to have "Dynamic" aspect. Verbs used to refer to states of affairs are said to have "Stative" aspect, and verbs used to refer to a habit or recurrent event type are said to have "Iterative" aspect.

Whether the syntactic contrast between the simple and progressive forms is better analysed through a complex tense system or through parallel tense and aspect systems is a question I am happy to leave to the linguists. Either way, a *semantic* notion of aspect is useful to identify contrasts in the temporal quality of an action or process, even when that is not signified by the syntax of the verbal group. The semantic notion is particularly useful for constructions involving modal auxiliary

verbs. Such constructions display no syntactic aspectual contrasts, since the modals themselves lack non-finite and participle forms (no **to can*, **canning* or **canned*[7]), and the main verb typically appears in a non-inflected form. In a modal auxiliary verb group, the modal itself has Stative aspect (Coates *loc. cit.*), but the main verb can display Dynamic, Stative or Iterative aspect, as illustrated by the literary examples using CAN in **Table 2.1**:

Although Coates correctly treats semantic aspect as a feature of the main verb in a modal verbal group, it seems no less appropriate to attribute that same aspect to the verbal group as a whole. The invariably Stative aspect of the modal auxiliary reflects a relationship internal to the verbal group, whereas the aspect of the main verb reflects the relationship of that verbal group to its subject and predicate. In what follows, I shall therefore treat aspect as a property of verbal groups as a whole.

Aspect type	**Example**	**Source**
Dynamic	For I **can** here disarm thee with this stick And make thy weapon drop.	Shakespeare: *The Tempest*, Act 1, Scene II
Stative	Then with the losers let it sympathize, For nothing **can** seem foul to those that win.	Shakespeare: *King Henry IV Part 1*, Act 5. Scene I.
Iterative	Betty **can** skip.	"Playmates" - The Victorian Reader, First Book, (1952, page 12)

Table 2.1

[7] Although *could* is the simple past tense of CAN, it is not its past participle. "*I have could." is not a sentence in English.

2.1.6 Semantic Categories of Modality

Apart from the epistemic/root distinction, modal verbs are used in English to express a variety of different relationships between a proposition or sentence and the world. A simple, non-modal verb group allows the speaker of a declarative sentence to assert that some fact obtains in the world. As we have seen in Section 2.1.3, a modal verbal group allows the speaker to assert either the possibility that some fact obtains, or the fact that some state of affairs is possible. But modality also deals with concepts other than possibility.

Von Wright offers a logician's analysis of modality that has found favour among linguists. He distinguishes four principal "*modi*" of modality, which he tabulates (Von Wright 1951: 3) as follows:

Alethic	*Epistemic*	*Deontic*	*Existential*
necessary	verified	obligatory	universal
possible		permitted	existing
contingent	undecided	indifferent	
impossible	falsified	forbidden	empty

On von Wright's taxonomy, alethic modality resembles epistemic modality, in that they both relate to the status of the proposition expressed by the verbal complement. Deontic modality deals with obligations imposed on agents, rather than natural constraints arising from the external world. Deontic modality is not relevant to the metaphysical description of free will I shall develop, but its mere existence in language provides indirect support for my claim that humans are capable of exercising free will. It would be pointless to oblige or forbid an agent to do some act, if whether that agent performs or refrains from performing the act were determined irrespective of the agent's knowledge of the obligation or prohibition. For that reason, I suggest that free will is presupposed in the often-cited maxim "ought implies can", where the "ought" expresses a deontic modality, and the "can" expresses the alethic modality relevant to free will, whose semantic nature will be made more precise in the sections to follow.

2.1.7 Coates' Empirical Studies

Linguist Jennifer Coates (1983) offers a semantics of English modal auxiliaries, based on actual written and spoken texts. Her data come from two sources. The written examples are taken from the Lancaster Corpus, a collection of printed texts containing approximately one million words gathered from fifteen genre categories. A further half million words of spoken material and some unprinted written material such as private letters and diaries are taken from the corpus of the Survey of English Usage assembled by University College London. From these data, Coates analyses in particular[8] the occurrences of ten modal auxiliaries, MUST, SHOULD, OUGHT, MAY, MIGHT, CAN, COULD, WOULD, WILL and SHALL, according to their semantic uses.

Although COULD and MIGHT are morphologically identical to *could* and *might*, the respective past tense forms of CAN and MAY, Coates treats them as separate modals, since they perform different semantic roles. Distinguishing between *could* as the preterite form of CAN and *could* as the preterite form of COULD is important in the free will debate, as will be seen later in this chapter.

Coates does not follow von Wright's taxonomy. In particular, rather than treating epistemic modality as an alternative to deontic, alethic and existential, she recognises that the epistemic/root distinction cuts across several other categories. From her data, she identifies nine major semantic categories of modality realised by the modal auxiliaries, namely Obligation, Inference, Possibility, Ability, Permission, Volition, Prediction, Hypothesis and what she calls "Quasi-subjunctive". Each of these categories can be realised by several of the modal auxiliaries, and each modal can realise several of the semantic categories. The relationships are shown graphically as in **Figure 2.1**, taken from Coates (1983: 26), where double lines signify the statistically most frequent use, single lines signify the next most frequent use, and broken lines signify infrequent uses.

---

[8] Her analysis also deals with quasi-modals such as HAVE TO, BE GOING TO, BE ABLE TO, BE BOUND TO. Of these, BE ABLE TO covers similar ground to CAN, which, rather than Coates' COULD, I claim is the modal relevant to an analysis of free action.

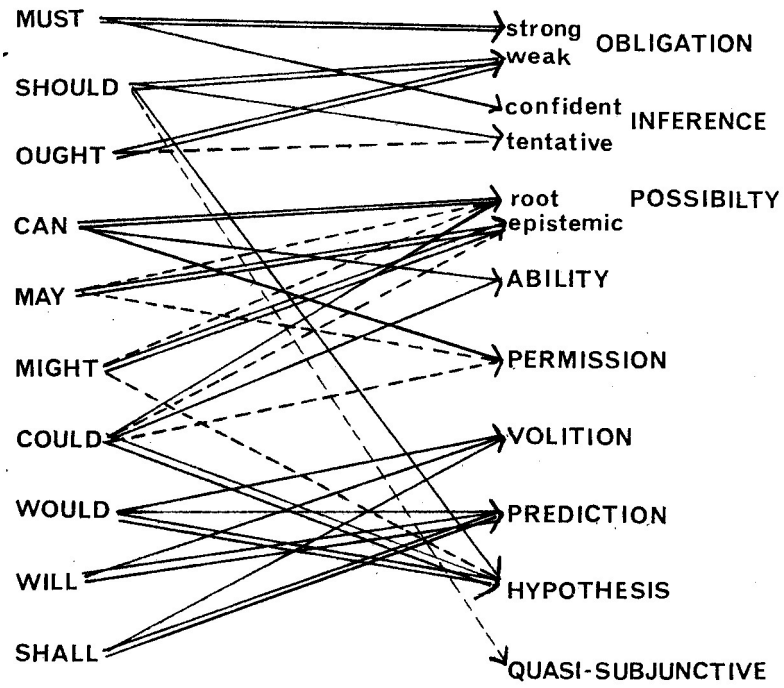

FIGURE 2.1

Von Wright's "deontic" modality covers much of the ground of Coates' Obligation, and Permission. (Root) Possibility and Ability largely correspond to von Wright's "alethic" modality. Von Wright's "existential" modality finds no place in Coates' taxonomy, but that is understandable. Von Wright's essay is an essay in modal logic, not in empirical linguistics. The relations describable through "existential" modality, though important to logicians, are not the subject of everyday discourse[9]. Also, those relations are more naturally expressed by quantifiers like "some" and "all", rather than by modal auxiliaries.

Of the ten modal auxiliaries studied by Coates, I shall consider only two, namely CAN (including its preterite form *could*), and COULD. In fact, I shall argue that free action is correctly analysed in terms of CAN, and that where *could* correctly appears in the analysis of a past instance of metaphysically free action, that word is used as the preterite form of CAN, and not as a form of a semantically separate modal verb COULD.

---

[9] From a total of 500 written texts, the Lancaster Corpus includes only 80 texts from the genre "Learned (including scientific)". Few if any of those would have been texts about logic. And to include a disproportionate number of philosophical or logical texts in any survey of linguistic usage would be unhelpful, to say the least.

As shown in Figure 2.1, CAN and COULD between them range over the semantic categories of Permission, Possibility, Ability and Hypothesis. CAN most commonly realises Root (though never Epistemic) Possibility[10], to a lesser extent Ability, and infrequently Permission. COULD most commonly realises Hypothesis, to a lesser extent Ability and Root Possibility, and infrequently Permission and Epistemic Possibility. Illustrative examples of each of those uses, from published sources, are shown in **Table 2.2**. (The terms "*realis*" and "*irrealis*" in the second column are explained in Section 2.1.7, below.)

2.1.8 The *Realis-Irrealis* Distinction

Another useful semantic distinction, explicit in the morphology of Spanish though not in that of English, is between *realis* and *irrealis* uses of a verb. In each of the above examples using the modal *can,* the proposition expressed is asserted as being true. Similarly, the examples of *could* expressing Ability, Root Possibility and Permission all assert the propositions in which the modal construction occurs. Such uses are termed *realis*, as they purport to describe reality. In contrast, the examples of *could* expressing Hypothesis and Epistemic Possibility refrain from asserting the propositions in which the modal construction occurs. The Hypothetical context asserts a counterfactual, in which the antecedent is denied, but neither the consequent nor its negation is asserted. Epistemic Possibility contexts characteristically withhold assertion of the proposition they express. Although expression of epistemic possibility is never inconsistent with the truth of the proposition expressed, use of the modal form in circumstances where the speaker is able to assert the bare proposition would contravene Grice's maxim of quantity (Grice 1975). Such uses are termed *irrealis*, as they treat the situation or state of affairs described as unreal.

---

[10] According to Coates, (*loc. cit.*:. 85), "CAN is the only modal auxiliary where we do not find the Root-Epistemic distinction".

Verb	Modality expressed	Example	Source
CAN	Root Possibility	Not all the water in the rough rude sea **Can** wash the balm off from an anointed king	Shakespeare, Richard II, Act 3 Scene II
CAN	Ability	I **Can** Jump Puddles	Marshall (1955)
CAN	Permission	Because I've told you before, oh, You **can**'t do that.	Lennon/McCartney (1964) A Hard Day's Night Album
COULD	Hypothesis (Counterfactual) (*irrealis*)	But that I am forbid To tell the secrets of my prison-house, I **could** a tale unfold …	Shakespeare, Hamlet, Act 1 Scene V.
COULD	Ability (past time)	Poor Alice! It was as much as she **could** do, lying down on one side, to look through into the garden with one eye …	Lewis Carroll (1865) Alice's Adventures in Wonderland, Chapter 2
COULD	Root Possibility (past time)	[T]he cosmic microwave background (CMB)… is believed to have originated … 300,000 years after the Big Bang, when …electrons and protons began to combine to form atoms, and radiation **could** propagate freely.	George F.R. Ellis (2002)"Cosmology: Maintaining the standard" Nature 416, pp. 132-133
COULD	Permission (past time)	First he got the Earth ready; then he got the Sea ready; and then he told all the Animals that they **could** come out and play.	Rudyard Kipling Just So Stories: The Crab that Played with the Sea
COULD	Epistemic Possibility (*irrealis*)	The thing was too easy, therefore it **could** not be. The hole would be empty.	Agatha Christie The Secret Adversary, Chapter 20

TABLE 2.2

Although Coates treats COULD and CAN as semantically distinct modal verbs, *could* is syntactically a form of CAN, and is treated as such in traditional grammar. Thus the Oxford English Dictionary has only one entry for *could*, namely:

> pa. tense (and *obs*. and *dial*. pa. pple.) of CAN *v*., q.v. Also *spec*. in ellipt. phr. ***could be***: it could be (that); it is possible; your suggestion may be correct.

□

In the following analysis, I shall therefore not treat *could* as a separate modal, but as the inflected preterite form of CAN. The preterite form (*could* used alone) encodes *either* past time (as in Ability, Root Possibility and Permission) *or* present (or timeless) *irrealis*, (as in Hypothesis and Epistemic Possibility), but not both. The past time examples in Table 2.2 are all *realis*. Past time *irrealis* is expressible, but not simply by combining the preterite *could* with an uninflected verbal complement. To express past time *irrealis*, it is necessary to use constructions such as *could have done* or *could have propagated,* where the past time sense is carried not by the modal *could*, but by the past participle form of the verbal complement. Such a construction is found in the phrase "*could have done* otherwise", so prominent in the philosophical free will debate, to be discussed in Section 2.2.5.

Although it is possible to find paradigmatic examples of each of Permission and Ability, Coates' analysis shows a gradient of meanings linking Permission to Possibility, and Possibility to Ability. The gradient from Permission to Possibility reflects the degree to which action is restricted by interpersonal obligations or social rules, rather than external circumstances. The gradient from Possibility to Ability reflects the degree to which action is restricted by circumstances external or internal to the agent. These gradients are illustrated by the examples in Table 2.3, taken from Coates' data.

The ordering of these examples is due to Coates. Examples 1 to 5 show what she calls the "Gradient of Restriction", from what is pretty clearly the speaker's granting of permission to what is possible under the laws of chemistry. Examples 6 to 11 illustrate Coates' "Gradient of Inherency", from what is made possible by factors external to the subject through to what is within the subject's abilities. Intermediate examples are open to differing interpretations. Example 3, for instance, might be

Permission	**1**	You can start the revels now.
	2	Poppy …can't drive because she hasn't got any insurance.
	3	[T]here are three answers they can give.
	4	[W]e can't expect him to leave his customers.
	5	How, then, can I help …impose a ban in which I do not believe?
Possibility	**6**	Salts can easily be separated from the solid residue by dissolving them.
	7	Well I think there is a place where I can get a cheap kettle.
	8	These young assistants …can give the pupils valuable practice in understanding and speaking the foreign language.
	9	Every believer can be a faithful distributor of the gospel.
	10	I can type only very slowly because I am quite a beginner.
Ability	**11**	It is now getting quite difficult to find choirboys old enough to behave in church who can still sing treble.

TABLE 2.3

taken as a pure possibility case if the question to which the answers may be given were "Is x greater than, less than or equal to y?". Coates' ranking would be appropriate for a multiple choice test where the examiner offers three choices. In real life, context could resolve such ambiguities, and likely Coates had access to the contexts in which the actual utterance was made, to justify her classification.

Free action, if it is ever manifested, is a capacity of agents. Believers in free agency say that a "free" agent can choose what to do in circumstances where an unfree "agent" could not. If such a distinction is meaningful, it is a distinction among

agents, and reflects something inherent in the agents. Thus, CAN='Ability' is at least one important sense of CAN for the analysis of free action, if not the only one.

According to Coates' analysis, there are three defining characteristics of CAN='Ability' , namely:

> (a) the subject is animate and has agentive function; (b) the verb denotes action/activity; (c) the possibility of the action is determined by inherent properties of the subject. (Coates (*ibid*.: 89)

Another characteristic shared by all the clear examples of CAN='Ability' in the corpora studied by Coates is that the verbal complement (or non-modal part of the verbal group) has Iterative aspect[11]. In contrast, clear examples of CAN='Possibility' show Dynamic aspect in their main predication.

The co-occurrence of CAN='Ability' with Dynamic aspect in the main predication is not precluded by Coates' schema, but no examples are to be found in her data. In Section 2.2, I shall argue that the metaphysical analysis of free action employs CAN='Ability' with Dynamic aspect of the verbal complement, but the absence of that combination from Coates' data is not at all surprising. As noted in Section 2.1.7, the texts included in Coates' data were drawn from a diverse range of genres, and those texts are unlikely to have included any philosophical discussions of free will. Outside philosophical discussions, constructions combining CAN='Ability' with Dynamic aspect of the verbal complement are understandably rare. In everyday discourse, (and for that matter in scientific texts), if we have occasion to assert that *A* can do *X*, it is because we want to deny that there are circumstances which would falsify that assertion. And the circumstances that might prevent *A* from doing *X* on any particular occasion seem to fall into two main classes. Either *A* lacks some general capacity or ability as demonstrated by his failure to do *A* on similar occasions when he has tried (CAN='Ability' with Iterative aspect), or circumstances external to

---

[11] Here, Coates distinguishes uses of CAN with a main verb of perception or mental process as in "I can see you" or "I can remember". In such uses, the main verb has Stative aspect, and the factuality of the unmodulated verb is implicit. If "I can see you", then I do see you. According to Coates (*ibid*., p. 90), such constructions "substitut[e] for "the unacceptable (*sic*.) progressive form **I am seeing*, etc." As reported by Crystal (1995), however, such progressive forms are perfectly acceptable in other dialects of English, such as those spoken in India.

A prevent his accomplishment of *X* when he tries on that particular occasion (CAN='Possibility' with Dynamic aspect). In ordinary or scientific discourse, there is rarely cause to deny that an agent who has demonstrated the ability to *X* on similar occasions and is not prevented from doing *X* on this occasion can do *X* on this occasion. Indeed, denials that an agent has a proven lack of ability to do *X* and that the agent is presently prevented from doing *X* by external circumstances allow the inference that the agent can do *X* on this occasion, only on the tacit assumption that the agent otherwise has the ability on this occasion. To make that assumption explicit would require CAN='Ability' with Dynamic aspect, but outside philosophical contexts, it is rarely if ever made explicit. Such freedom is taken for granted[12].

2.2 Analysing Free Action

2.2.1 The Concept of Metaphysical Free Will

I claim that human agents, and probably other animals, have "free will" in the sense that some of the things we do are "up to us". Furthermore, although it is not essential to my position, it seems plausible that in one natural sense of the word, every "act" of an agent is to some degree "up to" the agent. For reasons discussed in Section 1.4, any event over which a potential agent has no degree of control is arguably not an act of that agent, but at most an involuntary movement of some part of the potential agent's body.

---

[12] The folk view is also reflected in the criminal law, where persons' responsibility for their actions must be assessed. If it is found in a particular case that a person had no choice, or could not have done other than he or she did, the person should be acquitted. (Corrado 1991).

The Criminal Codes of Queensland and Western Australia, purporting to codify the common law, expressly exclude a person from criminal responsibility for an act or omission:

> … done or made under such circumstances of sudden or extraordinary emergency that an ordinary person possessing ordinary power of self-control could not reasonably be expected to act otherwise (Queensland 1899, s. 25).

Being limited to "circumstances of sudden or extraordinary emergency", the provision has rarely been invoked. But that very limitation reflects the view that being able to act otherwise is the normal condition of human beings in their engagement with the world.

The characterisation and individuation of "acts" is a complex topic. Intuitively, we might say an act is some event or complex of events which involve a person or some other conscious entity as agent. In a sentence reporting the act, the agent would appear as (logical) subject[13], rather than as object, but the metaphysical relationship between agent and act is far from clear.

When I speak of "acts" in this thesis, I have in mind certain events or causally-linked chains of events, typically including bodily movements of a conscious being, the agent. I would also include mental acts, such as performing mental arithmetic. Although these may not involve overt bodily movement, I assume they involve some physical changes in the neuronal state of the agent's brain. Typically the agent is human, but I would not exclude higher forms of animal life, nor the possibility of embodied extra-terrestrial beings.

As I use the word, "act" may refer to some relatively simple event such as raising an arm, or some complex of events such as climbing a mountain or writing a symphony. Complex acts of the latter kind may be recognised as composed of many smaller acts, some of which may be causally related to each other, and others of which may be linked only by the purpose the agent sets out to achieve. What characterises these events as acts is the existence of an agent and the fact that the agent seems to stand in some kind of causal relation to the event or to certain of the events in the complex.

Whether the agent stands at the head of the causal chain is controversial. I shall later claim that any act, properly described, originates with the agent, but as I use the term in this section, "act" would not exclude something performed by an agent who is himself or herself caused to perform it.

Goldman (1970) develops a comprehensive *Theory of Action,* according to which all acts "bottom out" with some overt bodily movement. My use of the term differs from Goldman's, as I would also include events internal to the agent's body,

---

[13] I refer here to logical subject, rather than grammatical subject. If the sentence is in the "active" voice, logical and grammatical subject coincide, but in passive voice constructions, the agent, if mentioned at all, becomes the indirect object. Systemic functional grammar avoids the ambiguity in "subject" by using the term "Actor" for the logical subject of clauses relating to what it terms "material processes". (Halliday and Matthiessen (2004: 53ff.) Notably the terminologies of both traditional and functional grammar reflect the intuitive conception of action.

such as voluntary muscular contractions and the firing of neurons, particularly where they are (partial) causes of overt bodily movements that qualify as acts.

In this sense of the word "act", I would argue that all acts are to some degree "up to us" when we eliminate all the different ways in which an act can fail to be free. (Section 1.4) But to support my thesis I need only an existential claim, namely that some events in the world are the result of choices by agents. Such choices are what I call exercises of free will, and the events resulting from such choices are what I call free actions. I claim to be one agent that sometimes exercises free will. I very much doubt that I am atypical of my species, but even if I were the only being endowed with free will, that would suffice for my existential claim.

In much of what follows, I shall speak in the first person because I believe I have privileged access to my own thought processes, but even if Ryle (1949) is right, I have no less access to my own mind than to that of anyone else. To sustain my claim that free actions sometimes occur, it suffices to establish that at least some things I "do" are to some extent "up to me". The reader is invited to consider, by analogy, whether some actions are "up to" him or her.

Let us take as an example my claim that it is now "up to me" whether I raise my arm within some specifiable time interval in the near future. In making that claim, I seem to be asserting the conjunction of the following two propositions, each of which can be taken as modified by an identical temporal adverbial phrase such as "within the next five seconds":

1. I *can* raise my arm.

2. I *can* refrain from raising my arm.

More will be said shortly about the function of *can* in the above two sentences. At this stage, I merely observe that the two uses of the word *can* are intended to be grammatically and semantically identical.

To say (without an auxiliary verb such as CAN) that "I raise my arm" implies that an objectively describable and observable physical event occurs, namely that my arm rises. But saying "I raise my arm" implies more than saying "My arm rises". There are circumstances in which I would assert the second, but not the first. Much

has been written by philosophers about the relationship between pairs of events such as raising arms and arms rising (Wittgenstein (1953, I 621)[14], Hornsby (1980a , 1980b), O'Shaughnessy (1973)), but the raising of an arm at least entails an arm's rising, and that suffices for the present discussion. To say that "I can raise my arm" implies that I have the ability, power or capacity to bring it about or ensure that my arm rises.

To say that I can refrain from raising my arm is not just to deny that I have the power to raise my arm. If it were, sentence 2. would merely be the negation of sentence 1., and their conjunction would be a contradiction in the form "*p* & ~*p*". The statue of Venus de Milo lacks the ability to raise its arm, but we would never say that it "can refrain" from doing so. To say that I can refrain from raising my arm requires *at least* that I have the ability, power or capacity *not* to bring it about or ensure that my arm rises. If, but only if, the truth of Sentence 1. is assumed, that may be a sufficient paraphrase of Sentence 2. If we contemplate an impossible (for me) action, such as levitating, the correlate of Sentence 1. is false, and it would also be inappropriate to say that I have the ability, power or capacity *not* to bring it about or ensure that my body levitates. There are lots of things I cannot do, but the only things it seems appropriate to assert I have the *power* not to do are those it is at least possible that I might do. Although the idea of a power not to do something need not be incoherent, the paraphrase is at best infelicitous. Powers are more naturally conceptualised with affirmative objects. The claim on some occasion that I can refrain from raising my arm is thus better expressed as the claim that on that occasion I have the power or ability to *ensure that it does not rise*[15].

To say it is up to me whether I raise my arm therefore seems to entail a conjunction of the following two propositions:

---

[14] What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?

[15] On some occasions when I claim to exercise my ability to ensure that my arm does not rise, it may happen that my arm rises without my bringing it about. Someone may grab it, or I may suffer some involuntary muscular spasm. Some Frankfurtian demon (Frankfurt 1969) may have intercepted and interfered with my neural signals. But when I (and I submit other normal users of English) say we can refrain from raising our arm, we normally expect that if we do not raise our arm, it will not rise. In the circumstances where our arm is grabbed or moved by an involuntary spasm or a Frankfurtian demon, we would assent to the statement that our arm rose, but would still deny that we raised it. On those surprising occasions when I claim the power to ensure my arm does not rise but it rises anyway, I shall have spoken falsely. I shall have failed on such occasions and in that respect to exercise my free will.

3. I *can* bring it about or ensure that my arm rises.

AND

4. I *can* bring it about or ensure that my arm does not rise,

where in each case, the word *can* may be paraphrased as "have the ability, power or capacity to …".

More generally, for some agent *A* to have free will requires that for some event *E*, the following conjunction be true:

5. *A can* bring it about or ensure that *E* occurs.

AND

6. *A can* bring it about or ensure that *E* does not occur.

If this conjunction turns out to be false, either on a particular occasion or in general, then it is false to claim, either on that occasion or in general, that *E* is "up to *A*". If it is false on some particular occasion, that shows that *A* like all humans is fallible. If it is false on all occasions and for all agents *A* and events *E*, there would be no free will in the sense described. Nevertheless, I claim the conjunction of 5. and 6. captures an important concept of free will, whether or not that concept is ever instantiated. In what follows, I shall refer to that concept of free will as "metaphysical free will".

2.2.2 Free Will in the Past

The analysis of a free action in Section 2.2.1 is expressed in the present tense. I choose the present tense for two reasons. My first reason is that as agents, we experience free will while we are exercising it. Although many of our actions take place with minimal deliberation, we agents are able on some occasions to deliberate and to experience what appears to be an exercise of free will in choosing what we shall do, and giving effect to that choice. After such an event, we may recall having exercised free will, and we may report on it using the past tense. Nevertheless, our most immediate evidence for having free will comes to us in and through its exercise.

My second reason for analysing free will in the present tense arises from the structure of the English language. Because the same preterite form *could* is used to express both past *realis* and *irrealis* verbal meanings, ambiguities are more likely to arise in the past tense than in the present tense, where the *realis can* contrasts with the *irrealis could*.

Bearing that in mind, the sentences 5. and 6. of Section 2.2.1 can readily be recast in the past tense. For some agent *A* to have had free will requires that for some past event *E*, the following conjunction is true:

7. *A could* bring it about or ensure that *E* occurred.

AND

8. *A could* bring it about or ensure that *E* did not occur.

The essential point to note is that the preterite *could* in each of 7. and 8. denotes simple past time, and conveys a *realis* meaning. Each conjunct asserts that the agent actually had a certain capacity at a past time. More will be said about this in Section 2.2.4.

2.2.3 The Semantic Characteristics of CAN and COULD

The *can*s and *could*s appearing respectively in sentences 5., 6., 7. and 8. differ from each other only in tense. As a group, but in contrast to many other *can*s and *could*s found in the philosophical free will debate, they share all of the following three characteristics.

Modality type = 'Ability' (rather than Possibility)	To say that "*A can* bring it about or ensure that *E* occurs", in the sense of 5. above, is to make a claim about the individual *A*. It ascribes to *A* some ability, power or capacity, rather than merely deny there is some external circumstance obstructing *A*'s performance[16].
Dynamic Aspect (rather than Iterative)	In the sense of 5., above, to say that "*A can* bring it about or ensure that *E* occurs" refers to single, potential token occurrence of *E*, rather than merely to assert that *A* has capacity to perform events of the type *E*, as demonstrated by past performances of events of that type.
Realis. (rather than *irrealis*)	In the present-tense, *can* is never *irrealis*[17]. In the sense of past tense sentences like 7. and 8. above, *could* is *realis*. The example makes a factual claim about *A*'s capacity to do *E* at the time to which it refers.

2.2.4 Retrospective knowledge

According to my analysis, the conjunction of 7. and 8. reports an actual (*realis*) state of affairs that existed at a past moment shortly before *A* exercised his free will. At that time, *A* would have spoken the truth if he had said "I can bring it about or ensure that *E* occurs", and he would have spoken the truth if he had said "I can bring it about or ensure that *E* does not occur".

---

[16] For the conjunction of 7. and 8. to be true on a particular occasion, it must also be true that there are no external circumstances preventing or necessitating the occurrence of *E*. But to say that it is "up to *A*" is to say more than that. It is to attribute to *A* the choice of whether or not *E* occurs.

[17] Arguably, *can* can be *irrealis* in future contexts, such as "If I get to the pub before it closes, I *can* have a beer". This *can* – paraphrasable as *shall be able to* – describes a possible future state of affairs that may never be realised, and therefore seems to be *irrealis*. However, the example illustrates CAN = 'Possibility' and not CAN = 'Ability'. Parallel examples involving CAN = 'Ability' seem artificial at best. Perhaps a paralysed man could say "If I get back the use of my legs I *can* walk", but in such a context, "*shall be able to*" sounds more natural than *can*. Happily, no analysis of free will I have seen has regard to future abilities, and future tense complications can be ignored in this chapter.

After the exercise of free will, we may well know whether *E* took place or not. Even if we don't know, there will be a fact that either *E* occurred or *E* did not occur. Let us assume that we now know that *E* occurred. Given that knowledge, it would be natural to say:

9. *A could* have brought it about or ensured that *E* did not occur.

In sentence 9., the modal verb form *could* is most naturally understood as past tense, rather than hypothetical. Although the speaker of 9. knows that *A* in fact did not bring it about or ensure that *E* did not occur, the verbal group as a whole remains *realis*, because the speaker asserts that *A* actually had the ability to perform *E* at the relevant time. The truth of 9. is consistent with the truth of the truth of both 7. and 8., but the utterance of 9. carries the further implication that *E* occurred. In circumstances where the speaker knows that *E* occurred, it would not be *false* to say:

10. *A could* have brought it about or ensured that *E* occurred,

but the utterance of that sentence by a speaker knowing that *E* occurred would violate Grice's (1975) cooperative principle (maxim of quantity).

2.2.5 "Could have done Otherwise"

Finally, in a conversational context predicated on the speaker's and hearer's shared knowledge that *A* did *E*, it would be natural to assert

11. *A could* have done *otherwise*.

In such an utterance, the adverb "otherwise" substitutes for some phrase such as "other than *E*", or "other than he did do", reflecting the speaker and hearer's shared knowledge that *A* did *E*.

In addition to sentence 11., we may construct a different sentence using the "otherwise" locution, but conveying the sense of sentence 8., uttered in the knowledge

that *A* did *E*. In such a context, we might say something like this: "We now know that *A* did *E* at time *T*. If *A* acted freely on that occasion, then at some moment prior to *T*, it was within *A*'s power to do *E*, and also at that moment, it was within *A*'s power to do other than *E",* or in other words:

> 12. (At that past time,) *A could* do otherwise (than that which he subsequently did, namely *E*).

I acknowledge that sentence 12. (even shorn of its parenthetical elaborations) is less commonly heard than sentence 11. Sentence 12. is likely to be uttered only by someone explicitly trying to elucidate the nature of a free action. Nonetheless, it is grammatically sound, and I submit, meaningful. It asserts unambiguously that, at the relevant past time, *A* actually had the power to do something other than that which he subsequently did.

Unlike sentence 12., sentence 11. is ambiguous. It could mean the same as sentence 12., and such a meaning was intended in the way sentence 11. was constructed. But the same words, "*A* could have done otherwise" are statistically more likely to be used in a hypothetical or counterfactual sense:

> (If he were not compelled to do *E*,) *A could* have done otherwise,

or in an epistemic sense:

> (For all I know,) *A could* have done otherwise.

In each of those senses, the modal *could* is *irrealis*. In such uses, the speaker either denies or expressly withholds assertion that *A* had the ability to do other than he did, in the circumstances actually obtaining. According to Coates' data (Coates 1983: 25), the most common use of COULD overall is *hypothetical,* and outside discussions of metaphysical free will, it is difficult to conceive of contexts in which a *realis* meaning of *could* would be intended in the phrase "could have done otherwise". But within discussions of free will, the phrase is often used, sometimes with a *realis* meaning

intended. The ambiguity of the modal form has led many theorists astray, as will be seen in Chapter 3.

2.3 The Reality of Metaphysical Free Will

2.3.1 Linguistic Evidence

In Section 2.1.6, I mentioned the maxim "ought implies can". In the light of Section 2.2.3, I can now say that the "can" implied is (like all instances of *can*) *realis*, and it expresses modality type ='Ability', with dynamic aspect. Human society depends to a large extent on notions of obligation amongst its members, and thus tacitly assumes that its members have, at least on some occasions, the capacities of the type expressed by that modality type.

Not only in our relations with others, but in the numerous acts of deliberation in which we engage, I suggest that all rational humans, whatever their theoretical commitments, conduct themselves as if they have the ability to give effect to the results of their deliberations. Though some philosophers seriously contend that it suffices that people act *as if* they have freedom in such a metaphysical sense, I suggest that people in general simply *believe* they can and do make free choices. Evidence for the generality of that belief is provided by the ubiquity of linguistic terms describing such choices, throughout human languages.

The categories of Roget's Thesaurus show how deeply the common belief in free action is embedded in the English language. In the Introduction to his original edition, Mark Peter Roget explains his intention to provide a semantic classification of words and phrases of the English language. Taking his guidance from "the principle … employed in the various departments of Natural History", Roget identifies six primary Classes or Categories, respectively dealing with (1) ABSTRACT RELATIONS, (2) SPACE, (3) the MATERIAL WORLD, (4) the INTELLECT, (5) the EXERCISE OF VOLITION and (6) SENTIENT AND MORAL POWERS. (Kirkpatrick 1998: xvi ff.) The fifth category takes up about twenty-two percent of the Thesaurus, and he explains it as follows:

> 5 The fifth class includes the ideas derived from the exercise of volition; embracing the phenomena and results of our Voluntary and Active Powers, such as choice, Intention, Utility, Action, Antagonism, Authority, Compact, Property, &c.

Although his Thesaurus is confined to the English language, Roget claims "the principles of its construction are universally applicable to all languages, whether living or dead". I believe that is a broadly plausible claim. Although the work of Benjamin Lee Whorf shows there may be disparities between the languages of radically different cultures about abstract concepts such as the nature of time, I am not aware of any culture so radically different from our own as to lack a semantic category for volitional concepts.

2.3.2 Metaphysical Free Will and Natural Selection

I can no more disprove that my choices are an illusion than I can disprove I am a brain in a vat. But independently of the mere subjective experience of acting freely, there is a further compelling argument that humans and other sentient beings have a genuine capacity to exercise choices which shape their environment. Notwithstanding Dennett's (1991) "explanation", consciousness remains a mystery to many of us. Yet I believe its existence is not something anyone can deny. Whatever its nature, it is a phenomenon that seems to be exhibited by some complex, living entities, and is lacking in simpler entities.

Among scientists and philosophers, it is generally agreed that the species of living organisms that now exhibit consciousness evolved from simpler entities by a process of natural selection. Creationists may disagree, but I suspect few of those who claim free will is an illusion are creationists. If the intervention of consciousness in shaping our future as humans were merely epiphenomenal, such consciousness would make no contribution to the fitness of an organism or species, but would be a costly and superfluous trait, unlikely to have evolved[18].

---

[18] My reliance on evolutionary theory should not be interpreted as inconsistent with my position that mental events and properties lie outside the scope of contemporary physics. As will be

This argument is by no means original. The significance of consciousness in evolution was discussed less than twenty years after the publication of Darwin's *On the Origin of Species* by William James. James poses the questions: "Of what use to a nervous system is a superadded consciousness? Can a brain which has it function better that a brain without it?" (James 1879: 4) According to him, "[c]onsciousness consists in the comparison of [simultaneous possibilities] with each other, the selection of some, and the suppression of the rest by the reinforcing and inhibiting agency of Attention" (*ibid*.: 13). He surmises that the selective process benefits a conscious being over an unconscious one.

A similar point is made by the physicist, Arthur Compton (1935) in a series of lectures published as *The Freedom of Man*. After explaining that quantum mechanics shows the world not to be deterministic, he observes, with reference to Nietzsche:

> A position frequently held by those who assume a deterministic world is that consciousness is so coördinated with the brain that whenever the physiological conditions require a course of action we are conscious of choosing to perform that action. Our purposes are "effective" and we feel "free," since what we choose to do is done; yet we are in no way responsible for our actions, because our wishes themselves are the necessary outcome of our past history and are not subject to our control.
>
> The most cogent objection that I find to this hypothesis is the complete uselessness of consciousness to the organism with which it is associated if the view is correct. If it is really ineffective, why should such an extraordinary thing as mind have evolved in association with certain types of organisms? It is an almost universal rule that in the evolutionary process those characters which are of value to a species persist and are accentuated, whereas those that are useless become vestigial or disappear. We find that among the higher animals the course of evolution has brought consciousness to an ever higher level of development. Why should this occur if consciousness were of no value to the life of the animal, and how could it be of value to the

---

discussed in Chapter 5, I think the best hope of accounting for free will is within an expanded ontology, encompassing entities or properties, not all of which are physical (in the sense of today's physics), but all of which are "natural", in the sense of being part of the reality that constitutes "nature". I take "natural" selection to involve selection by and within the totality of nature, whereas physicalists presumably would equate it with "physical selection".

Such a broad view of "natural" leaves no room for "supernatural" entities, or at least relegates them to the realms of fiction, so like the hardest of physicalists, I place no reliance on supernatural entities.

> animal if it were not capable of affecting its course of action? (Compton 1935: 54-6)

It is true that evolutionary processes can favour the development of some traits that themselves make no contribution to fitness, but come along "for free" as a by-product of some other favourable feature. Gould and Lewontin (1979) have emphasised that possibility with an architectural analogy. The dome of St. Mark's Cathedral is supported by four rounded arches. Among the most impressive features of the building its four decorated spandrels: triangular spaces defined by the exterior curves of mutually adjacent arches. As Gould and Lewontin point out, spandrels are "necessary by-products of mounting a dome on rounded arches". They serve no structural function. Similarly, evolution can favour traits which themselves make no contribution to the fitness of an organism or species, but occur as necessary by-products of some other feature which does.

Gould and Lewontin's target is what they call the Adaptationist Programme, according to which theorists were wont to offer hypotheses to explain the selection of individual traits in an organism. Apart from the difficulty of individuating such traits, their point is that not all traits favoured by evolution confer a fitness advantage, and they recommend comparison of organisms as integrated wholes. While they plausibly challenge a number of adaptationist theories, they do not offer a biological analogue of the spandrel.

The suggestion that consciousness may be an epiphenomenal by-product of evolution is anticipated and rejected by Compton:

> The reply is sometimes made that consciousness may be a by-product, similar to the lactic acid secreted by tiring muscles, which though useless may be an inescapable accompaniment of the functioning of the brain. Perhaps as good an answer as any to this objection is the statement made by Huxley in his *Elementary Physiology,* "How is it that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin when Aladdin rubbed the lamp". There is, in other words, absolutely no reason to expect a non-physical thing as consciousness to be the by-product of a physicochemical nerve current. (Compton 1935: 56)

It is implausible that consciousness and metaphysical free will are by-products of physical traits favoured by natural selection. Any feature that emerges as a by-

product of some other feature or combination of features must surely be of the same kind as the features that give rise to them. Thus in Compton's example, a chemical substance is secreted as a by-product of the useful chemical processes that take place in muscles. In Gould and Lewontin's example, spandrels are geometric features that arise from the construction of a structure with certain geometric constraints. Though almost all chemical processes have by-products[19], and features of shape may follow as by-products of other features of shape, it is hard to accept that an epiphenomenal consciousness would emerge as a by-product of the physical structure of a brain, nervous system or other biological entity.

2.4 The Importance of Metaphysical Free Will

In Section 1.3, I mentioned that a major part of the philosophical literature on free will is motivated by moral questions. If, as widely though not universally believed, humans are to be held accountable only for those actions they "freely" perform, society must have some agreed criteria by which an action is judged to be "free". The concept of metaphysical free will was intentionally developed in Section 2.2 independently of any moral consequences of any action freely performed, however, and any connection between that concept and the types of freedom presupposed in moral judgements needs to be demonstrated.

I claim that many, if not most, of the actions resulting from free choices by human agents have no moral consequences whatsoever, or at least none that can be discerned or predicted, and could thus be influenced by moral considerations. Whether or not I now raise my hand while sitting alone at the keyboard is, I claim, up to me and I can discern no way in which my doing so or refraining from doing so will affect the interests of any other being. In contrast, a judge who by raising his hand can prevent the death of another human being and refrains from doing so (as in van

---

[19] The distinction between products and by-products of a chemical reaction seems to be relative to some interest or other. Carbon dioxide and an excess of yeast are produced in the brewing of beer, and are also useful as an industrial gas and as the precursor to Vegemite®. It seems natural to call them "by-products" because we think of the place they are produced primarily as a brewery.

Inwagen's thought experiment, Section 3.2) makes a choice with major moral consequences. Such a judge may well deliberate at great length, and his eventual choice will be influenced – even in *some* sense "determined" – by his regard for the moral consequences. Yet as physical systems, the judge and I are similar, and it would be strange if the causal processes leading up to our hand raising or non-hand raising behaviour were fundamentally different, merely because of the societal convention by which his action is accorded some meaning which mine lacks.

I claim that metaphysical free will is an important concept because it plays a part in all free actions. It is important in my present project because I believe its incompatibility with causal closure – to be demonstrated in Chapter 3 – tells us something important about the universe. But it should also be important to those whose prime concern is with moral issues, because metaphysical freedom is involved in all choices, including those whose consequences give rise to responsibility, praise and blame. In our weightiest deliberations, when as moral agents we deliberate and agonise over what we should and should not do on a particular occasion, I say we do so only because we take it for granted in some sense or other that the moment will come when we will have to choose between alternatives, either or all of which are in some sense within our power.

Not all deliberation leads to an exercise of free choice, however. It is a sad fact about humans that we are apt to reflect and retrospectively deliberate about what we should have done at some past time, even though we know such retrospective deliberation does not allow us a retrospective freedom to choose. "If only I had bought Telstra shares at $3.75, I could have recovered some of my losses!" A person engaging in such fruitless deliberation knows it is too late to revise the lamented choice, but the remorse experienced depends on the belief that there was a time at which a different choice would have been effective. While we may feel great regret about the 2004 tsunami, and wish it had not occurred, no rational person wishes they had prevented its occurrence, because the occurrence of the tsunami was never within any person's power to prevent.

From my perspective, whether people have metaphysical free will is an interesting question about the world, principally because the fact of its possession provides us with information about the causal structure of the universe; namely that not all events have their probabilities determined by prior physical states and events in

combination with physical laws. But for those whose philosophical interest is motivated primarily by moral questions such as the justification of praise and blame, the existence of metaphysical free will should also be an important issue, if the very processes of deliberation in which we engage when deciding what we should do (or judging what an agent should have done), presuppose that, at the end of the deliberating process, we shall have (or the agent had) the ability to give effect to the conclusion of the deliberations.

While few would deny that people act *as if* they have free will, some like Wegner (2002) maintain that the feeling of acting under the idea of freedom in the metaphysical sense is no more than an illusion. Unless there is some independent reason to believe metaphysical freedom is impossible, such as a proof that the physical domain is causally closed, I submit that the "best" explanation for the feeling of freedom is that agents in fact have a capacity to exercise real choices. In chapter 4, I shall argue that physics itself gives no reason to believe that the physical domain is causally closed.

2.5 Wegner's Illusion

Although I claim in Section 2.3 that the feeling that people have free will is best explained by the fact that they really do have free will in the metaphysical sense of Section 2.2, Daniel Wegner has forcefully argued that the feeling that we consciously will our voluntary actions is an illusion:

> It is an illusion in the sense that *the experience of consciously willing an action is not a direct indication that the conscious thought has caused the action.* (Wegner 2002: 2)

Wegner draws a distinction between "the *experience* of consciously willing an action" and "the *causation* of the action by the person's conscious mind" . These, he claims are entirely distinct.

Although Wegner's topic is "conscious will" rather than "free will", his claim that the former is an illusion equally applies to any experiences that the conscious will is acting freely. He briefly dismisses the conflict of free will and determinism as

based on a "false dichotomy", and lampoons the "robogeeks" and "bad scientists" who participate in the debate:

> The … clash fails on both sides because free will is a feeling whereas determinism is a process. They are incommensurable. (*ibid*.:322)

His own concept of free will admits no middle ground between determinism and randomness, and claims that free will theorists confuse having free will with experiencing free will. That dichotomy reflects his main dichotomy between the experience of free will and the unperceived processes that cause actions:

> [H]aving free will [and] experiencing free will … are vastly different. Having it doesn't happen, whereas experiencing it goes on pretty much constantly. (*ibid*. 324 *fn*.)

Wegner presents an impressive body of empirical evidence to show that the experience of conscious will and the actual processes of mental causation are different things. His evidence shows there are circumstances in which people seem to others to be acting consciously, yet deny having the experience. Also, there are cases where people wrongly believe they are causing some event or other. Yet I would deny that those circumstances show that the experience of conscious will *never* reflects a genuine causal process. On the contrary, those exceptional cases are worthy of notice only because they are a departure from our normal experiences. Wegner is quite explicit in drawing the inference. In a response to several critics of his 2002 book who made a similar point, he retorts:

> Several of the commentators …argue that the listing of exceptions to the efficacy of conscious will does not invalidate conscious will overall, and that the book's project is therefore in error.
> This doesn't make sense to me. If someone has a theory that the earth orbits around the sun, for example, and it turns out that on Wednesday, February 4, 2004, it briefly orbited around a Wal-Mart store in Duluth, I'd say that pretty much shoots the sun theory. Exceptions do not prove rules, we all know – they *invalidate* them.
> … This logic requires that we draw an important inference. From *The feeling of conscious will can be mistaken,* we must conclude that *the feeling of conscious will is never correct*. I agree that on its face, this seems extreme – But there is a deeply important reason for making this strong inference: Imagine the horrible kludge we would

> have to create to accommodate a *partly* valid experience of conscious will in any mental system we might envision underlying human behavior. In essence, we would need to create a double system – one to produce willed acts in which the feeling was authentic, and another to produce acts in which the feeling is disconnected from the action and did not reflect how the action has occurred. (Wegner 2004: 682-3)

I find that argument quite unconvincing. One might argue in the same way that all visual perceptions are illusions because some of them are, and we don't want a "horrible kludge" of theories to explain the veridical ones as well as the illusory ones. It is true that exceptions invalidate universally quantified rules, but they do not invalidate existential claims. A black swan invalidates "All swans are white", but an albino leopard does not invalidate the claim that "Some (or even most) leopards have spots".

Wegner claims people have developed the experience of conscious (or free) action because they frequently foresee their actions before they happen. From a Humean constant conjunction of experiences of will and the actions foreseen comes the illusion that the will causes the actions, whereas both the will and the action have a common cause in some "intricate set of physical and mental processes" (2002: 27) of which the agent is not conscious. As Wegner admits (*ibid.*: 318), the experience of will is epiphenomenal, but he claims it is no less important for that.

On Wegner's account, the illusion of conscious will is important, because it is only through a sense or "emotion" of authorship that we can keep track of our causal role in the world around us. For that purpose, the conjunction of our experiences of will with the acts that are ours is sufficient. And according to Wegner, the illusion of conscious will is also sufficient to satisfy our needs as social beings:

> Conscious will is particularly useful, then, as a guide to our selves. It tells us what events around us seem to be attributable to our authorship. This allows us to develop a sense of who we are and are not. It also allows us to set aside our achievements from the things that we cannot do. And perhaps most important for the sake of the operation of society, the sense of conscious will also allows us to maintain the sense of responsibility for our actions that serves as a basis for morality. (*ibid.*: 328)

However much the illusion of free will might give us the emotion of authorship, and bestow a sense of responsibility, it does not account for the common belief that people

sometimes have real choices. From an agent's point of view, there is a real difference between the experience that one is about to sneeze (involuntarily) and that one is about to raise one's hand where one might have chosen not to. In the latter case, it is not just that we sometimes anticipate our hand rising and other times not, but that it is sometimes "up to us" what we do. If the feeling that some things like hand raising are "up to us" is itself an illusion, I do not believe Wegner's theory accounts for that illusion. Here, the distinction between "conscious will" and "free will" becomes important.

In Section 2.3.2, I also claimed that unless human agents have metaphysical free will in a robust sense, there is no plausible explanation of how consciousness and any epiphenomenal sense of acting freely has evolved under natural selection. Wegner certainly gives good reasons why it is advantageous to organisms to have the experience of conscious will. The sense of authorship is something that humans value, and would, I suspect, continue to value even if it ceased to be veridical. But unless that sense of authorship played some causal role in the interaction of our ancestors with their environment, I do not see how it or the consciousness that values it could have evolved in the first place.

Close to the end of *The Illusion of Conscious Will*, Wegner makes the following claim:

> Illusory or not, conscious will is the person's guide to his or her own moral responsibility for action. If you think you willed an act, your ownership of the act is established in your own mind. You feel guilty if the act is bad, and worthy if the act is good. … Guilt…, pride…, and the other moral emotions… would not grip us at all if we didn't feel we had willed our actions. Our view of ourselves would be impervious to what we had done, whether good or bad, and memory for the emotional content of our actions would not guide us in making moral choices for the future. (*ibid.*: 341)

I can agree with the sentiments expressed in that passage, but if conscious will is no more than an illusion, I wonder how Wegner would account for the "moral choices" to which he refers or for the "us" that is guided in making them. If "us" refers to the consciousness that has merely epiphenomenal experiences, it cannot choose future events. If the "us" refers to the unknowable "intricate set of physical and mental

processes" that cause our actions, Wegner gives no account of how those processes are affected by the experiences of the conscious will.

2.6 Conclusions

I submit that the best explanation for the common feeling that people can make choices which affect their physical environment is that we possess metaphysical free will as characterised in Section 2.2.1. Such metaphysical free will is tacitly presupposed in the ordinary processes of deliberation, and if it were merely an illusion, the faculty of consciousness could not have evolved through natural selection.

Although not all exercises of metaphysical free will have moral consequences, all moral choices are made on the tacit assumption that the agent has the metaphysical capacity to give effect to the choice when made. Thus metaphysical free will is not a mere curiosity, but a capacity underlying any system of ethics.

Apart from its importance to ethics, the acknowledgement of a robust capacity for metaphysical free will leads to an important conclusion about the causal structure of the world. In the next chapter, I shall demonstrate that metaphysical free will in the sense defined is incompatible with the principle of causal closure of the physical domain. That leads to the conclusion that some physical events, namely those giving effect to the metaphysically free choices of conscious agents, have causes that are not themselves physical. As such a conclusion is in some sense dualistic and thus at odds with current orthodoxy, I shall defend the position at some length in the remaining chapters.

CHAPTER 3

INCOMPATIBILISM

3.1 Introduction

In Chapter 2, I defined a concept of metaphysical free will and distinguished that concept from other concepts that are also of philosophical interest. I argued not only that metaphysical free will is a capacity displayed by normal human agents on innumerable occasions, but that its possession is pre-supposed in the exercise of other concepts of human freedom of interest to philosophers.

In this chapter, I shall claim that metaphysical free will as defined in Chapter 2 is incompatible with a widely accepted metaphysical doctrine, namely that the physical domain is causally closed in the sense that all physical events, unless they are uncaused, have at least their probability of occurrence determined by other physical events and states of affairs. After examining van Inwagen's Consequence Argument for the incompatibility of free will and determinism, I shall offer additional arguments to show that free action in the metaphysical sense just mentioned is not only incompatible with physical determinism, but also with causal closure of the physical domain.

Using the semantic metalanguage developed in Chapter 2, I shall then examine the writings of philosophers who have argued that "free will" – in some sense or other – is compatible with determinism or causal closure of the physical domain. By analysing the modal language used by those philosophers, I shall show that the notions of "free will" claimed by them to be compatible with determinism or physical causal closure differ significantly from the metaphysical concept of "free action"

whose incompatibility with causal closure of the physical domain I claim to demonstrate.

3.2 Compatibilism and the Burden of Proof

A widely favoured view among philosophers at present is that, contrary to what the common folk might think, "free will" is not incompatible with "determinism". I agree to this extent: the compatibility of "freedom" with "determinism" is indeed contrary to what the common folk think. Perhaps that is a point on which I can speak with some authority, having spent half a century as one of the common folk before becoming exposed to the views of the philosophical community. I suspect few writers on this subject have had the opportunity to reflect for so long in isolation. I suggest at least that people outside the philosophical community take for granted that they have free will in the sense that some things are "up to them". People who are not philosophers probably don't reflect much on determinism, but I suggest that if asked, most lay people would reject the idea that the things that are "up to them" were nevertheless made inevitable by circumstances beyond their control.

But regardless of what the common folk may believe, many philosophers, for well considered reasons, are compatibilists. And given the widespread acceptance of one or another form of compatibilism among philosophers, I accept that any denial of compatibilism (or to put it more positively any assertion of a particular incompatibilism) needs to be defended.

Lycan (2003) issues the following challenge:

> Compatibilism, not just about free will but generally, on any topic, is the default. For any modal claim to the effect that some statement is a necessary truth, I would say that the burden of proof is on the claim's proponent. A theorist who maintains of something that is not obviously impossible that nonetheless that thing *is* impossible owes us an argument. And since entailment claims are claims of necessity and impossibility, the same applies to them. Anyone who insists that a sentence S1 entails another sentence S2 must defend that thesis if it is controversial. If I tell you that 'Pigs have wings' entails 'It snows every day in Chapel Hill,' you need not scramble to show how there might be a world in which the first was true but the second false;

> rather, you would rightly demand that I display the alleged modal connection.□ And of course the same goes for claims of incompatibility.
>
> The point is underscored, I think, if we understand necessity as truth in all possible worlds.□ The proponent of a necessity, impossibility, entailment or incompatibility claim is saying that *in no possible world whatever* does it occur that so-and-so.□That is a universal quantification. Given the richness and incredible variety of the pluriverse, such a statement cannot be accepted without argument save for the case of basic logical intuitions that virtually everyone shares.

In response to Lycan's challenge, I shall express my incompatibilist claim in the form of a necessary truth. So expressed, my claim is that in no possible world is it the case both that a person can act freely in the sense of Sentences 5. and 6. of Section 2.2.1, and that all future events in that world have their probability of occurrence fixed by the laws of nature and the state of that world at some time prior to that person's free action. I claim that is a necessary truth, because the incompatibility is inherent in the *concept* of acting freely as set out in Sentences 5. and 6.

In the context of the free will debate, compatibilism and incompatibilism are usually presented as contrasting theses about "free will" in some sense or other, and determinism. For example, in the *Oxford Companion to Philosophy*, we find a short entry that begins as follows:

> **Compatibilism and incompatibilism.** Compatibilism is a view about determinism and freedom that claims we are sometimes free and morally responsible *even though* all events are causally determined. Incompatibilism says that we cannot be free and responsible *if* determinism is true. … (Honderich 1995, italics added)

The apparent contrast between the italicised words "*even though"* and "*if*" in that entry seems to suggest that compatibilism endorses the truth of determinism while incompatibilism withholds judgement. I suspect that contrast is not intended by either the editor (a leading incompatibilist) or the author. Whatever the author intended, the broad consensus among physicists and probably metaphysicians is that some events at the sub-atomic level are inescapably and irreducibly indeterministic. Compatibilist philosophers are likely to concede there is indeterminism in sub-atomic systems, but to claim that such indeterminism is irrelevant to the question of free will. A common

approach is to argue for the compatibility of "free will" and determinism, and then to claim in a footnote that the argument can be tweaked, without losing its validity, to allow for the fact that it is only the *probabilities* of events that are determined. Lycan, *loc. cit.*, exemplifies just such a strategy:

> For the sake of argument, I shall assume that determinism is true. I myself believe that it is not true; but let us assume it, because my main thesis is the compatibility of freedom *with determinism*, and because if I am right …, indeterminism would not help anyway. Also, if no incompatibilist argument succeeds, there is no reason to think that indeterminism would make us any freer.

The incompatibilism thesis I wish to defend is that it is not possible for a person to be able to act freely in the sense of Sentences 5 and 6 above, if all future events, including the consequences of that person's actions, have their probability of occurrence fixed by the laws of nature and the state of that world at some time prior to the supposedly free action. My arguments in support of that thesis are developed from the "Consequence Argument", introduced in van Inwagen (1975), elaborated upon in van Inwagen (1983), and recently refined in van Inwagen (2002). My arguments differ from van Inwagen's in two respects. Firstly, they specify the sense of acting freely as the metaphysical sense of Section 2.2, and thereby avoid van Inwagen's problematic concept of "rendering a proposition false". Secondly, while van Inwagen argues only for the incompatibility of free will and determinism, the second of my arguments seeks to show that metaphysical free will is also incompatible with causal closure of the physical domain.

3.3 van Inwagen's Consequence Argument

The consequence argument is perhaps the best-known argument against compatibilism, and its most prolific defender is Peter van Inwagen. Van Inwagen calls the argument "obvious", but then explains:

> I don't mean it's obviously right; I mean it's one that should occur pretty quickly to any philosopher who asked himself what arguments could be found to support incompatibilism (van Inwagen 1983: 16).

Informally stated, the argument relies on the intuition that some of our actions are "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: 56)

In a more formal version of the argument, van Inwagen relies on the concept of an agent "rendering a proposition false". For him, each proposition is either true or false. He rejects any talk of propositions being true at particular times, but loosely speaking, we have to understand that any proposition has the same truth value at all times. At least no proposition is true at some time and false at some other time. Of course, identical sentences, uttered at two different times, might denote two different propositions with different truth values.

In van Inwagen's terminology, an agent has the power to render a true proposition false if and only if it is within that agent's power to do something which, if he were to do it, that proposition would be false. If the agent exercises the said power, then the proposition, which in the actual world is (timelessly) true would be (timelessly) false. In Section 3.5, I shall consider whether a more plausible account can be given if we allow propositions about the future to lack a truth value, though such an account would not be acceptable to van Inwagen.

3.3.1 van Inwagen's Formal Argument

Van Inwagen formalises his consequence argument by describing a thought experiment in which we are asked to imagine a judge, *J*, who by raising his hand at a certain time *T* might have prevented the execution of criminal, but who refrained from doing so. Van Inwagen argues that *J* could not have raised his hand at *T* if determinism is true. In the argument set out below, "T_0" denotes an arbitrarily

chosen instant of time earlier than *J*'s birth, "P_0" denotes a proposition that expresses the state of the world at T_0, "*P*" denotes a proposition that expresses the state of the world at *T*, and "*L*" denotes the conjunction into a single proposition of all the laws of nature. Van Inwagen claims that the seventh proposition follows deductively from the other six:

(1) If determinism is true, then the conjunction of P_0 and *L* entails *P*

(2) It is not possible that *J* have raised his hand at *T* and *P* be true

(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_0 and *L* entails *P*, then *J* could have rendered the conjunction of P_0 and *L* false

(5) If *J* could have rendered the conjunction of P_0 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*. (van Inwagen 1983: 70)

This argument seems to be on the right track, though the concept of rendering a proposition false leads to a difficulty noticed by Lewis, and discussed in Section 3.4 below. The consequent of premise (1) is exactly what van Inwagen understands by determinism. Premise (2) is a logical truth, since *P* includes a statement of the fact that *J* did not raise his hand at *T*. The consequent of premise (3) seems to follow from premise (2) in that if *J* had raised his hand at *T*, the state of the world would have included his hand being in a raised position, contrary to the specification of *P*. Premise (4) instantiates the plausible principle that if one can render a proposition false, one can thereby render false any further proposition which entails that proposition. Premise (5) instantiates the principle that if one can render false the conjunction of two propositions, one of which concerns only states of affairs pertaining before one's birth, one must do so by rendering false the other proposition.

Premise (6) states that one cannot render a law of nature false, which is plausible, according to a normal concept of laws of nature.

3.3.2 The Grammar of "Could"

Because van Inwagen constructs his thought experiment in the past, he needs to use past tense verb forms to describe it. Perhaps anticipating potential confusion between what I call *irrealis* and *realis* uses of *could*, he expressly warns the reader:

> Note that all the conditionals that occur in (1)-(7) are material conditionals: the 'could have' that occurs in them is merely the past indicative of 'can'. (van Inwagen 1983: 70)

I find that statement puzzling. In traditional grammar, the past indicative of CAN is *could*. In the indicative mood, the auxiliaries *can* and *could* respectively are followed by the uninflected root form of the main verb, as in the pair of sentences "Betty *can skip* today." and "Betty *could skip* yesterday.". It is difficult to see how *could have* is a past indicative of CAN. In a sentence such as "Before they closed the pub, we *could have* lunch in the beer garden", *could* alone functions as the past tense of the auxiliary CAN, and *have* functions as the main verb. In clauses such as van Inwagen's "J could have raised his hand at *T*", *could* is not followed by an uninflected main verb, but by a further auxiliary *have* and by a past participle *raised*. The assertion that "J could have raised his hand at *T*" carries not only the implication that in some sense *J* had the ability to raise his hand, but also the implication that he did not do so. And indeed, that implication is consistent with the story told by van Inwagen in sentences (1) to (7) above.

In traditional grammar, " …'Mood' refers to verbal inflections or to syntactic contrasts that (1) denote by formal opposition the relations between one verb in the sentence and another verb structure, and (2) express a notional contrast that supposedly indicates the attitude of the speaker or writer toward the action or state of affairs expressed by the verb" (Harsh 1968: 12-13). The indicative mood contrasts with the imperative mood (not relevant here) and the subjunctive mood. The contrast between the indicative and subjunctive moods is between whether "… the speaker or

writer considers a syntactic structure as stating a fact … or as expressing non-fact or modification of fact" (*ibid.*). The syntactic category indicative therefore corresponds to the semantic category of *realis*, while subjunctive corresponds to *irrealis*.

Many modern grammarians would deny that contemporary English has a subjunctive mood, because there are no inflections used uniquely for expressing *irrealis*. But undeniably, different inflections (or uninflected forms) are used to create contrasts between *realis* and *irrealis* clauses, and the term "subjunctive" is often used to refer to the verb in those *irrealis* forms, as in "If I *were* you …", and " …move that the motion *be* put". In the former case, the preterite form is used to refer to a counterfactual state of affairs, and in the latter, the uninflected infinitive form is used for a desired state of affairs yet to be realised.

For my own analysis, I prefer the semantic terms *realis* and *irrealis*, but when van Inwagen refers to the "past indicative", I take him to be denying that the *could*s in his argument are in the subjunctive mood, or in my terminology, *irrealis*.

To say simply that "J *could have raised* his hand" implies the unreality of the hand rising, but not the unreality of the capacity or ability. There is a potential for misunderstanding the sense of *could* in van Inwagen's argument, because the *could have* form is also used in the consequent of a counterfactual. Thus to say that "If *J* were not paralysed, he could have raised his arm at *T*" implies not only that *J* did not raise his arm, but also that he lacked the capacity or ability to raise his arm. In such a sentence, the *could* is *irrealis*, (or subjunctive in van Inwagen's terminology), but no such clauses appear in van Inwagen's argument. One way of avoiding such a misunderstanding would be to paraphrase "could (not) have [past participle]" as "had (or did not have) the ability [infinitive]. The argument then becomes:

(1) If determinism is true, then the conjunction of P_0 and L entails P

(2) It is not possible that *J* have raised his hand at *T* and *P* be true

(3) If (2) is true, then if *J* had the ability to raise his hand at *T*, *J* had the ability to render *P* false

(4) If *J* had the ability to render *P* false, and if the conjunction of P_0 and L entails P, then *J* had the ability to render the conjunction of P_0 and L false

(5) If *J* had the ability to render the conjunction of P_0 and *L* false, then *J* had the ability to render *L* false

(6) J did not have the ability to render *L* false

(7) If determinism is true, *J* did not have the ability to raise his hand at *T*.

3.4 Lewis's Challenge to van Inwagen

David Lewis (1986b) argues against van Inwagen's argument in the form of Section 3.3.1, claiming there is an equivocation between premises (5) and (6) in the phrase "could have rendered false". According to Lewis's analysis, "*J* could have rendered *L* false" can be understood either weakly or strongly as follows:

(Weak Thesis) *J* was able to do something such that, if he did it, a law would be broken.

(Strong Thesis) *J* was able to break a law.

If the argument is to succeed, "*J* could have rendered *L* false" must be given the same interpretation in each of premises (5) and (6). According to Lewis, on the weak interpretation, premise (6) is false, while on the strong interpretation, premise (5) is false. Lewis himself accepts the weak thesis. The judge could have raised his hand, and *if* he had done so, some law would have been broken, but the judge could not have broken any law.

Lewis's weak thesis seems to apply only in cases where the subject does not, in fact, do what he is able to do. For the sake of generality, let *A* be something that *J* was able to do such that if he did *A*, a law would be broken. The weak thesis guarantees that there is such an *A*. If we add the proposition that *J* did *A*, then we also have the proposition that a law was broken. But as Lewis himself affirms, (*ibid.*: 292) "… to say that anything [was] both a law and broken … is a contradiction in terms". And it seems the only way to avoid that contradiction is to deny our supposition that *J* did *A*.

So while the weak thesis gives a sense in which *J* could have raised his arm, that sense is only available in a case where *J* did not raise his arm. And that seems to exploit an asymmetry between things that could have been done and were done, and things that could have been done and were not done.

3.5 Facts about the future

Van Inwagen insists that every proposition is simply true or false, and refuses to relativise the truth value of propositions to particular times. According to van Inwagen, if phrases such as " … 'true at some particular moment', 'true at every moment', 'became true', 'remained true', 'is unchangeably true' and so on" (1983: 35) are meaningful at all, they must be understood as saying that the proposition that would be expressed by the uttering phrase at the time mentioned would be a true proposition.

Van Inwagen's concept of rendering a proposition false must be understood in the light of that insistence. To render a proposition false at time *T* does not mean that a proposition was true up to that time, and then became false thereafter. Undoubtedly, the same words uttered on two different occasions can refer to propositions of which one is true and the other false, but for van Inwagen, every proposition is timelessly true or timelessly false. If I freely raise my hand at time *T*, a consequence of that choice is that the proposition expressed by the (tenseless) sentence "My hand rises at *T*." is (timelessly) true. The truth of the proposition is a consequence of my free choice, and in that sense, I rendered it true.

In the light of these views, van Inwagen seems to be committed to saying that when an agent makes a choice between two actions, it is determinately true that he will perform one of those actions and not the other. That is tantamount to accepting an asymmetry between those things the agent is able to do and does, and those things an agent is able to do but does not. It is on such an asymmetry that Lewis's challenge to van Inwagen's argument relies.

I share van Inwagen's discomfort at the idea that true propositions can become false or *vice versa*, but I am not committed to the idea that all propositions have a truth value at every moment of time. In particular, I suggest it is not incoherent to say

that propositions about the future events *presently* have no truth value, although *timelessly* they are (the "are" being tenseless) either true or false.

Metaphysical free will, as described in sentence pairs 5. and 6. or 7. and 8. of Section 2.2 is crucially symmetrical. If an agent right now has a genuine choice whether or not to do *E*, the characterisation of that state of affairs must not depend on what that agent's choice turns out to be. If there is genuinely an open future – if an agent right now has equally the power to do *E* and the power to ensure the non-occurrence of *E* – there is, right now, no fact about whether *E* occurs. Of course there will come a time at which we could look back and see whether or not *E* occurred. At that time, the proposition that the agent did *E* will be either true or false. And from a God's-eye perspective outside time, that tenseless proposition that the agent does *E* has the same truth value. But that does not entail that the tenseless proposition has any truth value right now, and to insist otherwise would be to deny that the agent has metaphysical free will in the sense of Section 2.2.

The distinction I draw is between a proposition being timelessly true (or false) and that proposition being true at all times. I contend that at least some propositions about future events or states of affairs do not presently have truth values, though from a God's-eye perspective, outside time, they are either true or false.

I acknowledge that from a timeless perspective, there is a fact about whether I submit my thesis by 5. p.m. on 21st February 2006. The timeless perspective is unattainable, reminiscent of Archimedes' point or Nagel's (1986) "view from nowhere". McCall's (1994) branching model of the universe also needs to be described from a perspective outside time. As an aspiring metaphysician, I don't wish to renounce the right to talk about things from that perspective. Nevertheless, I deny that, as I write these words, the statement "I submit (tenseless) my thesis by 5. p.m. on 21st February 2006" is (present tense) either true or false.

From the timeless perspective, the sentence "I submit (tenseless) my thesis by 5. p.m. on 21st February 2006" has a truth value. I deny the inference from that to the sentence "I submit (tenseless) my thesis by 5. p.m. on 21st February 2006" having a truth value at the time I write these words. I think that inference unjustifiably equates timeless truth, or truth from outside time, with truth at all times. If *p* is true *at all times*, then *p* must be true now, but the timeless truth of *p* does not guarantee the truth of *p* at any given instant of time, nor quantified over all times. Timeless truth is truth

without time, and that is a different concept from truth *at each and every* time. Some truths (like mathematical and logical truths) may be timelessly true *and also* true at all times, but that is no reason to assume that the two concepts are the same.

In the next section, I offer some alternative arguments for the incompatibility of free will with causal closure, expressed in the present tense. These arguments rely crucially on the symmetrical concept of metaphysical free will as defined in Section 2.2. Because Lewis's challenge to van Inwagen relies on the asymmetry between what the agent is able to do and does, and what the agent is able to do but does not, I claim that a similar challenge would not defeat the arguments based on the symmetrical concept of metaphysical free will.

3.6 Incompatibility Arguments in the Present Tense

A way to avoid the ambiguity between *realis* and *irrealis* forms of *could*, and also to ensure there is no asymmetry that depends on what the agent in fact turned out to do, is to recast the consequence argument in the present tense. Some modification is required, however. If I act freely in the sense of 5. and 6. of Section 2.2, then even if my future free act is determined, it is presently impossible to state a proposition about what that act will be. Lewis, in his (1986b) criticism of van Inwagen, sets out the Consequence Argument in the first person. I shall follow his example, as I believe it is from our own first-person experience that agents derive our conception of acting freely.

I shall offer two versions of the present-tense argument. The first aims to show the incompatibility of free action and determinism, and the second aims to show the incompatibility of free action and causal closure of the physical domain. These arguments do not employ the concept of rendering a proposition false, but develop from the concept of acting freely as encapsulated in Sentences 5. and 6. of Section 2.2. Whether or not people actually *have* freedom in that sense is not at issue here. I seek only to establish that freedom in that sense would not be possible, firstly in a deterministic world, and secondly in any world in which the probability of future events is determined by the present state of the world and the laws of nature.

3.6.1 Incompatibility with Determinism

For me to be free – here and now – to raise my hand, requires both that it be true that I can in the next five seconds (say) raise my hand, and also that it be true that I can for the next five seconds refrain from raising my hand. Let "*P*" denote a proposition that expresses the present state of the world, let *T* be some instant in the near future, and let "*L*" denote the conjunction into a single proposition of all the laws of nature. On the assumption of metaphysical free will, there is presently no fact about whether or not I shall raise my hand, and we need to argue hypothetically from two alternative assumptions. An informal version of the present tense argument runs as follows:

> Suppose that I have a free choice (in the sense of Sentences 5. and 6. of Section 2.2) about whether or not I raise my hand. That is to say, right now:
>
> > I can bring it about or ensure that my hand rises by *T*.
> > AND
> > I can bring it about or ensure that my hand does not rise by *T*.
>
> Now suppose that the world is deterministic. Then it follows that the conjunction of *P* and *L* determines either that my hand will rise by *T* or that that my hand will not rise by *T*.
>
> Assume firstly that the conjunction of *P* and *L* determines that my hand does *not* rise by *T*. This means that the proposition that my hand will rise by *T* is false, and contradicts the assumption that I can ensure that my hand rises by *T*.
>
> Now assume that the conjunction of *P* and *L* determines that my hand *does* rise by *T*. This means that the proposition that my hand will rise by *T* is true, and contradicts the assumption that I can ensure that my hand does not rise by *T*.
>
> Whichever disjunct we assume, we have a contradiction of the conjunction required for me to be free to raise my hand. Thus, my freedom to raise my hand or not to do so is incompatible with determinism.

3.6.2 Formal Statement of the Argument Assuming Determinism

More formally, let *R* stand for the (timeless) proposition that my hand rises at some specific time *T* in the near future. And as before, let "*P*" denote a proposition that expresses the present state of the world and let "*L*" denote the conjunction into a single proposition of all the laws of nature. Then to say that it is presently up to me whether I raise my hand at *T* requires the truth of the following conjunction of two propositions, both indexed to the present moment:

1. I can ensure *R* and I can ensure ~*R*

If determinism is true, the present state of the world and the laws of nature determine either that my hand rises at *T* or that it does not rise at *T*. In other words:

2. *P* and *L* determine *R* or *P* and *L* determine ~*R*.

To claim that free action in the sense described above is incompatible with determinism is to claim that 1 and 2. cannot be true together. The following argument shows that these two propositions, taken together, lead to a contradiction:

1.	(I can ensure that *R*) & (I can ensure that ~*R*.)	Def. Freedom
2	(P and *L* determine *R*) v (P and *L* determine ~*R*.)	Def. Determinism
3	I can ensure that *R*	1 &E
4	I can ensure that ~ *R*.	1 &E
5	*P* and *L* determine *R*	Assumption
6	*R*	(see below)
7	~(I can ensure that ~*R*.)	(see below)
8	(I can ensure that ~*R*) & ~(I can ensure that ~*R*)	4, 7 &I

9	It is not the case that *P* and *L* determine *R*	5, 8 RAA
10	*P* and *L* determine ~*R*	2, 9 disjunctive syllogism
11	~*R*	(see below)
12	~(I can ensure that *R*)	(see below)
13	(I can ensure that *R*) & ~(I can ensure that *R*)	3, 12 &I

The respective propositions of lines 1 and 2 define freedom and determinism. Propositions 3 and 4 follow logically from proposition 1 by the rule of conjunction elimination. Line 5 assumes the first disjunct of proposition 2.

Proposition 6 is not a logical consequence of proposition 5, but follows from the concept of determinism. If the present state of the world and the laws of nature determine the truth (at all times) of the proposition *R*, then given that the present state of the world and the laws of nature obtain, *R* must be true at all times, including the present time.

Proposition 7 follows from proposition 6. It is impossible for both *R* and ~*R* to be true. So in any world in which *R* is presently true, ~*R* is *not* true, and in such a world it is not within my power, right now, to ensure that ~*R be* true.

Proposition 8, the conjunction of propositions 4 and 7, is a contradiction. Proposition 9 follows by *reductio ad absurdum*, being the negation of assumed proposition 5. Proposition 10 then follows from propositions 2 and 9 by the rule of disjunctive syllogism.

The inferences from proposition 10 to propositions 11 and 12 are identical to those from proposition 5 to propositions 6 and 7.

Proposition 13 is a contradiction formed by the conjunction of propositions 3 and 12. It follows as a consequence of propositions 1 and 2, and shows that at least one of those propositions must be false.

All of the inferences in the argument are logical, except the inference from 5 to 6 (and 10 to 11), and the inference from 6 to 7 (and 11 to 12). The inference from 5 to 6 (and 10 to 11) is unlikely to be contentious. If determinism is true, then whatever is determined to be true is true.

The inference from 6 to 7 (and 11 to 12) is in my opinion valid, but requires elucidation. The claim is that, *R* being a true proposition, it is not the case that I can ensure the truth of the false proposition ~*R*. To say in line 6 that *R* is a true proposition is not just to say it is timelessly true in the sense of Section 3.6. It is also true at all times, including the time expressed by the present tense in the argument. That is because line 6 is a consequence of line 5, which assumes one of the disjuncts of line 2 which in turn defines determinism.

The validity of the claim is ensured by the nature of the modality expressed by the verbal group "can ensure". The argument seeks to demonstrate the incompatibility of determinism with the metaphysical concept of freedom expressed by proposition 1, in which the modal verb *can* has a *realis* meaning, has Dynamic aspect, and expresses the modality type = 'Ability'. To say that I can ensure the truth of ~*R* is to make a factual claim about myself as an agent, here and now, given the world and the laws as they actually are. It is not to say that in some hypothetical circumstance or in some other world I or my counterpart would do something such that ~*R* would be true.

It is important to keep clear the tensed nature of the argument. The proposition in line 1 is a present-tense proposition, expressing the definition of metaphysical free will from Section 2.2. It expressly refers to the agent's capacity at a time when the tenseless proposition *R* has no truth value[20]. In contrast, the proposition in line 2 is a definition of determinism and as such assumes *R* has a truth value at all times. Similarly, *R* in line 6 has an assumed truth value, because it is derived from line 2 via the assumption in line 5.

Many authors have argued for the compatibility of some kind or other of freedom with determinism. And many of their arguments characterise freedom in terms employing the modal verbs CAN or COULD. No doubt there are meaningful and relevant conceptions of freedom which are compatible with determinism. But so far

---

[20] The rule of inference from 6 to 7 (and 11 to 12) in the above proof is available only because *R* (and ~*R*) in 6 (and 10) were assumed to have truth values. If *R* (and ~*R*) have truth values, then the negation of line 1 can be derived from the law of excluded middle as follows:

(i) ~*R* v *R*
(ii) ~(I can ensure that *R*) v ~(I can ensure that ~*R*.)
(iii) ~((I can ensure that *R*) & (I can ensure that ~*R*.))

But this merely shows that metaphysical free will is inconsistent with statements about the future such as *R* having a present truth value, which has already been acknowledged.

as I know, no previous author has characterised an agent's freedom in terms of a conjunction such as line 1 of the above proof, in which *can* expresses a *realis* modality with Dynamic aspect and modality type = 'Ability'. Certainly no previous author can (= 'Possibility') have demonstrated the compatibility of such freedom with determinism, since the two are incompatible, as shown by my argument.

There are many things that I, like all agents, can do but do not. Ensuring the truth of a false proposition is not one of those things. It is truly something I *do* not do, but it is also not something I *can* do, if the *can* is understood to express a *realis* modality with Dynamic aspect and modality type = 'Ability'. That kind of modality may be contrasted with that contained in David Lewis's (1986b: 293) ability "to do something such that if [he] did it a law would be broken". Lewis's ability there is contingent upon his not, in fact doing that thing. That contingency upon the non-performance of the contemplated action shows that the modality he expresses (through the phrase "I am able to") is not of modality type = 'Ability', although it does seem to be a *realis* modality with Dynamic aspect.

Given a choice between R and $\sim R$, the sense in which Lewis is able to do something (say R) which he in fact does not do is different from the sense in which he is able to do the thing he does in fact do ($\sim R$). In one case, he is bound (or determined) not to do the thing ($\sim R$). In the other case, he is bound (or determined) to do the thing. That asymmetry is not consistent with the metaphysical sense of freedom described in Sentences 5. and 6. of Section 2.2.1.

3.6.3 Incompatibility with Causal Closure

In an informal statement of the second version of the argument, P, L and T are defined as before.

> As in the first version, to say that I have a free choice (in the sense of Sentences 5. and 6. above) about whether or not I raise my hand means that, right now:

I can bring it about or ensure that my hand rises by *T*.

AND

I can bring it about or ensure that my hand does not rise by *T*.

Now suppose that the world is one in which the physical domain is causally closed. Then it follows that the conjunction of *P* and *L* determines the probability that my hand will rise by *T*.

Let the probability assigned by *P* and *L* to the proposition that my hand does *not* rise by *T* be *q*. If *q* = 1, the chance of my hand rising by *T* is infinitesimally small[21], which for all practical purposes contradicts the assumption that I can ensure that my hand rises by *T*.

Alternatively, if *q* < 1, there is a non-zero probability that my hand will rise by *T*, which contradicts the assumption that I *can ensure* that my hand does not rise by *T*. That contradiction is not just a latent one, possibly to be manifested at time *T*. It is a present one, existing at the time I utter sentences 5. and 6. For it to be true of me that I can ensure that my hand does not rise at *T*, (with *can* expressing *realis* modality with Dynamic aspect and modality type = 'Ability') it must be presently within my power to make it inevitable that it does not rise. If factors external to me, (*P* and *L*) determine a finite probability that the hand will rise, its not rising is not inevitable, and it is not within my power to make it inevitable.

It is true that on any single occasion, my hand may not rise. But our assumptions also ensure that on every such occasion there is a possibility that my hand does rise. Because that non-zero possibility is a fact about the universe, there is nothing I can do to *ensure* that my hand will not rise, even if it happens not to. If my hand did not rise, that would not be attributable to me, but to the conjunction of *P* and *L*.

The previous conclusion seems right to me. If I am free in the sense of Sentences 5. and 6., whether or not my hand rises is up to me. If the probability is fixed by *P* and

---

[21] Strictly speaking, it cannot be said that the proposition that my hand will rise is false. From any continuum, the probability of any single value being selected is zero, but that does not make it false that some value will be selected.

L, then whether or not my hand rises on a given occasion is *not* up to me. Still, an outside observer sees only that my hand rises or does not rise on any given occasion, and might object that my use of the concept "up to me" is unacceptably subjective. I therefore offer the following alternative argument.

> Assume on some occasion that the probability of my hand not rising, as fixed by *P* and *L*, is *q*. If $q = 0$ or $q = 1$, we can derive a contradiction as in the deterministic case, so let's assume *q* lies between 0 and 1. If I am free in the sense of Sentences 5. and 6., then I can ensure on this occasion that my hand does not rise. Let's assume that I so ensure. The failure of my hand to rise on this occasion is consistent with the finite probability, *q*, that it would not rise. But now let us consider a plurality of possible worlds with identical histories up to this moment. If *q* is determined by *P* and *L*, it is statistically likely that in some of those worlds the hand will rise, and in some it will not. As the number of worlds, *N*, increases, the so-called "law of large numbers" predicts, with increasing approximation to certainty, that the hand will rise in $N \times (1-q)$ worlds and will fail to rise in $N \times q$ of those worlds.
>
> Now let us assume that in half of those worlds I and my counterparts all ensure that my or their hand does not rise, and in the other half my counterparts ensure that it does. Either $q = 0.5 = (1 - q)$ or $q > 0.5$ or $(q - 1) > 0.5$. If $q \geq 0.5$, then as *N* increases, it becomes highly probable that in at least half of any set of *N*/2 worlds, the hand does not rise. Yet this cannot plausibly be true of the set of *N*/2 worlds in which I and my counterparts ensure that it does rise. But alternatively, if $q < 0.5$, as *N* increases, it becomes highly probable that in more than half of any set of *N*/2 worlds, the hand rises. And that cannot plausibly be true of the set of *N*/2 worlds in which I and my counterparts ensure that it does not rise[22].

---

[22] For any value of *q*, we have only to imagine a sufficiently large number of possible worlds to render the likelihood as small as we choose, and that number rises exponentially. When $q = 0.5$, and the number of worlds is 10, the probability that the hand fails to rise by chance in the five worlds in which I choose not to raise it and rises by chance in the five where I choose to raise it is less than 1 in a thousand. If the number of worlds is a hundred, the corresponding probability is less than 1 in 10^{30}. As *q* deviates further from 0.5, the probabilities become smaller. For $q = 0.25$ or 0.75 and $N = 10$, the probability is less than 1 in four thousand. For $q = 0.25$ or 0.75 and $N = 100$, the probability is less than 1 in 10^{36}. For $q = 0.1$ or 0.9 and $N = 10$,

The foregoing argument is symmetrical and makes no presupposition as to whether or not my hand rises by *T*. We could equally have assigned the probability *q* to the proposition that my hand *does* rise by *T*, and reached a similar result, *mutatis mutandis*.

3.7 Compatible Freedoms

3.7.1 Purpose of this Section

My conclusion that metaphysical free action is incompatible with both determinism and causal closure is at odds with the dominant views, at least among analytical philosophers, firstly that free will and determinism are compatible, and secondly that the kind of indeterminism that seems to be required by quantum mechanics does not affect the issue. I claim that arguments in favour of compatibilism, where they succeed, rely on different conceptions of "free will" or "freedom", and that any such compatible conceptions of freedom are insufficient to explain the experience of human agents that some of our actions are "up to us", irrespective of whether that experience is real or illusory.

In the remainder of this chapter, I examine a variety of compatibilist accounts, to show through the modal language used that they involve concepts of freedom significantly different from the metaphysical freedom developed in Section 2.2.

As mentioned in Section 1.3, a major part of the literature on the problems of free will is concerned with moral, rather than metaphysical questions. Many of the best-known papers on compatibilism are concerned primarily, if not exclusively, with reconciling physical determinism with society's practices of holding individuals accountable or responsible for at least some of their actions. It is not obvious[23] that metaphysical freedom as characterised in Section 2.2 is a necessary condition for the

---

the probability is less than 1 in a hundred thousand. For $q = 0.1$ or 0.9 and $N = 100$, the probability is less than 1 in 10^{52}

[23] Though, I would argue, true.

attribution of praise and blame, and many authors have identified and described other senses of freedom which fulfil that role.

Certainly, courts of law do not in practice require proof of metaphysical free will to convict defendants of criminal misdeeds, but I suggest that is because metaphysical free will is not called into doubt. In cases where a defendant claims to have been acting under duress, for example, the court will consider whether the threats of another party prevented the defendant from doing otherwise. But that question is only considered on the tacit assumption that if the agent had not been not under the alleged constraint, he could have done otherwise. Whether a defendant could have done otherwise only becomes an issue when some abnormal circumstances are alleged to have constrained the agent from the normal condition in which agents are assumed to act freely. It would be a rare case in which the defence rested solely on a claim that the defendant's actions had their probabilities determined by the state of the world before the agent was born, and the success of such a defence, if upheld on appeal, would destroy the criminal legal system.

Although I have argued in Section 2.4 that the kind of freedom recognised by society as necessary for moral responsibility *does* presuppose metaphysical freedom on the part of the agent, such a contention is not essential to my claim that metaphysical freedom is incompatible with determinism and causal closure. Quite apart from its connection with moral issues, metaphysical freedom is also of interest because its reality and incompatibility with causal closure tells us something interesting about the nature of the universe.

3.7.2 Conditional Interpretations of CAN and COULD

In his 1912 book, *Ethics*, G.E. Moore offers a consequentialist account of right and wrong actions. In the final chapter, he examines the concept of free will required for that account. His main claim is that attributions of right and wrong to an action " … depend upon what the agent *can do*" (Moore 1912: 122, italics in the original).

Moore is a compatibilist. He considers there is undeniable evidence that humans have free will, in the commonly understood sense that " … we sometimes *can*, in *some* sense, do what we don't do" (*ibid.*: 126), and is also unwilling to deny

determinism in the sense that " … in *one* sense of the word 'could', nothing ever *could* have happened except what did happen"." (*ibid.*: 130). But for Moore, these propositions employ different kinds of modality:

> [I]t is extremely doubtful whether Free Will is at all inconsistent with the principle that everything is caused. Whether it is or not, all depends on a very difficult question as to the meaning of the word 'could'. All that is certain about the matter is (1) that, if we have Free Will, it must be true, in *some* sense, that we sometimes *could* have done, what we did not do; and (2) that, if everything is caused, it must be true, in *some* sense, that we *never could* have done, what we did not do. What is very *un*certain, and what certainly needs to be investigated, is whether these two meanings of the word 'could' are the same. (*ibid.*: 130-1)

One sense in which people often claim that they 'could' have done something is to mean that they would (or should[24]) have done that thing if they had chosen (or decided, or wanted or willed) to do so. And Moore suggests that some such paraphrase is sufficient to justify the intuitions underlying his ethical theory, without denying the thesis of determinism.

Whether or not such a hypothetical sense of freedom is compatible with determinism, it should be clear that the metaphysical freedom described in Section 2.2 is not a hypothetical one. At issue in a hypothetical analysis of free action is whether and in what sense an agent *could* have done other than he in fact did. The question "*Could A* have done otherwise?" cannot arise until *A* has done some act with which "otherwise" provides a contrast. If "*A* could have done otherwise" is to be paraphrased as "*A would* have done otherwise if …", then whatever circumstances are offered after the word "if" will be counterfactual. The paraphrase will refer to *A*'s capacity in circumstances other than those that actually obtained, and the modality expressed in "could have done otherwise" will be *irrealis*, in the terminology of Section 2.1, rather than *realis* as required for metaphysical freedom. Indeed, no

---

[24] Moore uses 'should' rather than 'would' here as a first-person form of the conditional (*irrealis*) auxiliary, and not in any obligatory sense. Current Australian English uses 'would' for all persons. Gowers (Fowler1965, p. 713) describes this usage of 'should' as one in which "… the English of the English differs from the English of those who are not English". He wryly observes that, despite American influences, "…there are still [in 1965] Englishmen who are convinced that their *shall* and *will* endow their speech with a delicate precision that could not be attained without them and serve more important purposes than that of a race label".

analysis of "could have done otherwise" captures the concept of metaphysical freedom, for as explained in Section 2.2.5, the *could* is necessarily *irrealis*, whereas metaphysical freedom is properly described with a *realis* modality.

Although any proof that a hypothetical sense of freedom is compatible with either determinism or causal closure would not affect my claim that metaphysical freedom is incompatible, there remains a widespread view that a hypothetical paraphrase somehow disposes of the "free will problem". I shall therefore offer further observations on the merits of the hypothetical interpretation itself.

An objection that readily comes to mind is that any consideration of an agent's ability to do otherwise *if* he had so chosen seems to require that (in some sense of 'could') the agent *could have chosen* to do otherwise. If the second 'could' has the same sense as the first, a regression seems to follow. In the cited chapter, Moore anticipates such an objection but avoids the regression by treating the second *could* an epistemic, rather than hypothetical:

> It is therefore quite certain (1) that we often *should* have *acted* differently, if we had chosen to; (2) that similarly we often should have *chosen* differently, *if* we had chosen so to choose; and (3) that it was almost always *possible* that we should have chosen differently, in the sense that no man could know for certain that we should *not* so choose. All these three things are facts, and all of them quite consistent with the principle of causality. (Moore 1912: 136-7)

Quite apart from the semantic differences between epistemic modality and the Root ='Ability' modality required by metaphysical free will, I think such an explanation fails to capture any sense of free action. Surely whether an agent acts freely should be a question of fact. And questions of fact, unless they be facts about knowledge, should not depend on what is known.

The apparent regress in Moore's conditional paraphrase is also noted by Kane (1996) and Chisholm (1964). Chisholm argues that "You would have done otherwise if you had willed or chosen otherwise" entails "You could have done otherwise" only if "You could *also* have *willed* or *chosen* otherwise", and leads to a regress. Davidson (1973) claims that the regress stops after a few iterations, when active verbs such as "choose" can be replaced with dispositional verbs such as "want/desire/prefer". But in a deterministic or causally closed world, dispositions

would be determined just as much as actions, and I do not see how a compatibilist, defending a deterministic or causally closed world, can consistently leave a phrase such as "could have *desired* otherwise" unanalysed.

3.7.3 Austin's *Ifs and Cans*

Conditional" analyses of "can" and "could have" are criticised on grammatical grounds by J.L. Austin in his lecture "Ifs and Cans", presented to the British Academy in 1956. Austin summarises his response to Moore as follows:

> (a) 'I could have if I had chosen' does not mean the same as 'I should have if I had chosen'.
>
> - In neither of these expressions is the *if*-clause a 'normal conditional' clause, connecting antecedent to consequent as cause to effect.
> - To argue that *can* always requires an *if*-clause with it to complete the sense is totally different from arguing that *can*-sentences are always to be analysed into sentences containing *if*-clauses.
> - Neither *can* nor any other verb always requires a conditional *if*-clause after it: even 'could have', when a past indicative, does not require such a clause: and in 'I could have if I had chosen' the verb is in fact a past indicative, not a past subjunctive or conditional. (Austin 1961: 165)

Austin's distinction between the past indicative and past subjunctive uses reflects the semantic distinction between *realis* and *irrealis* uses of *could*, discussed in Section 2.2.5[25]. Austin draws on Latin grammar, where different inflected forms make the distinction clear:

---

[25] When Austin contrasts indicative and subjunctive uses of "could", he treats mood as a semantic concept. Following Lyons (1977, p. 746), Palmer (1990, pp. 11-12) treats grammatical "mood" as a syntactical concept, reflected through contrasting inflectional systems. In that strict sense, English makes minimal use of subjunctive mood. According to Palmer:

> [w]hat is sometimes identified as the subjunctive in English is either the past tense being used for unreality, especially in conditional sentences …, or the rather formal use of the simple, uninflected, form of the verb in subordinate clauses:
>
> If he came tomorrow …
> I suggest he come tomorrow.

> [I]t is natural to construe 'could have' as a past subjunctive or 'conditional', which is practically as much as to say that it needs a *conditional* clause with it. And of course it is quite true that 'could have' *may* be, and very often is, a past conditional: but it is also true that 'could have' may be and often is the *past (definite) indicative* of the verb can[26]. Sometimes 'I could have' is equivalent to the Latin 'Potui' and means 'I was in a position to': sometimes it is equivalent to the Latin 'Potuissem' and means 'I should have been in a position to'. …It is not so much that … 'could have' is ambiguous, as rather that two parts of the verb *can* take the same shape (*op. cit.*: 163)

In a famous footnote, Austin offers the following example of a context in which he claims a modal *could* is not to be understood conditionally:

> Consider the case where I miss a very short putt and kick myself because I could have holed it. It is not that I should have holed it if I had tried: I did try, and missed. It is not that I should have holed it if conditions had been different: that might of course be so, but *I am talking about conditions as they precisely were*, and asserting that I could have holed it. There is the rub. Nor does 'I can hole it this time' mean that I shall hole it this time if I try or if anything else: for I may try and miss, and yet not be convinced that I could not have done it; indeed , further experiments may confirm my belief that I could have done it that time although I did not. (Austin 1961: 166 fn., italics added.)

Here, the *could* in line 2 is not intended as hypothetical, since it expressly does not make a claim about what would have happened in different circumstances. Yet there seems to be something counterfactual in the statement, since Austin did not hole the putt, and the past *irrealis* form (*could have* with a past participle) reflects that.

As Palmer himself notes, this is "essentially a terminology question". If the terms "indicative" and "subjunctive" are to be reserved for purely morphological contrasts, the semantic contrast noted by Austin might more accurately be described as a contrast between *realis* and *irrealis*.

[26] *Pace* Austin, I think it would be more correct to identify *could,* rather than *could have* as the simple past form. *Could hav*e needs to be followed by a past participle like 'holed', while the paraphrase Austin offers, 'I was in a position to' needs to be followed by an uninflected infinitive like 'hole' or a compound form like 'have holed' in which the auxiliary *have* is itself an uninflected infinitive, and is not part of the modal auxiliary. Be that as it may, Austin explicates an important ambiguity between two uses of *could* followed by *have*.

Although Austin's locution resembles my description of metaphysical free will in its focus on actual circumstances, Austin is not reporting an exercise of free will. I think the difference becomes apparent when we reflect on how Austin would justify his assertion. Imagine him shaping up to the ball and saying "I can hole this". If we then asked him to justify that assertion, he would probably refer to his having performed similar acts in the past, and to the conditions surrounding him at the time. His statement, after the event, that he could have holed the putt is a perfectly natural one, and we would not want to accuse him of speaking untruthfully. He is making neither a metaphysical claim nor a hypothetical claim, but an epistemic claim, expressing a belief justified by historical and circumstantial knowledge. Such an epistemic interpretation is reflected in the past *irrealis* form of the verbal group.

If Austin the philosopher were to interrupt his golf game long enough to indulge our interest in metaphysics, he might entertain a different claim. To make a claim about metaphysical free will, he would have to assert not only "I can hole this", – this time with a *realis* dynamic modality of type ='Ability' – but also "I can refrain from holing this". The second of those statements could easily be true. He has only to pick up the ball and return to the clubhouse, but as things turn out in the original story, he would be speaking falsely by making the first assertion. As a prudent philosopher, he may decline to make the metaphysical claim, but if he does make it, he speaks falsely. He speaks falsely, because as the story turns out, he misses the putt. Therefore, he did not have the ability to bring it about or ensure that the ball go into the hole, as required in Sentence 5. of Section 2.2.1.

Does that mean he lacks free will on that occasion, or that we all lack free will on all occasions? Not at all. Austin does not hole the putt, but in his failed attempt, he does lots of other things. He *tries* to hole the putt. So when he falsely said "I can hole this.", he might truthfully have said "I can try to hole this.". If the notion of "trying" is physically suspect, we can consider the chain of physical events that occur when Austin tries to hole the putt. He hits the ball. He swings the club. He moves his arms. For at least one of these physical events (let's call it *E*), Austin would speak truthfully if he says, in the metaphysical sense, "I can do *E*". If there were no event *E* in that physical complex of events for which Austin could truthfully make that statement, I would be disinclined to describe the chain of events as Austin's act.

Although Austin mentions the short putt only in a short footnote, it has aroused the interest of other philosophers, including Honoré and Dennett, each of whose own views will be discussed below.

3.7.4 Honoré on 'Can' and 'Can't'

Honoré (1964) distinguishes a particular sense from a general sense of "can" and "could". To say an agent *A* can (general) do *E* means that in similar circumstances, if he tries, he will (usually) succeed. To say he can (particular) means that on this very occasion, if he tries, he will succeed. This latter analysis differs from hypothetical analyses such as Moore's, since the "if" statement is not counterfactual. Indeed, the "if" clause is not the antecedent of a conditional statement upon whose truth the consequent depends. Such non-conditional uses of "if" are common and innocuous in everyday dialogue. Austin (1956: 158) gives the example "There are biscuits on the sideboard if you want them". Honoré's example, "if he tries, he will succeed" is less innocuous, but also not a conditional on Honoré's account. It is rather a claim that "he normally succeeds in doing it when he tries". As Honoré explains it, "the 'if' clauses refer to the proof , not the truth, of the 'can' statements. Thus 'I can if I choose' may be expanded to read 'I can, and if I choose to do the action you will see that I can, because I shall succeed". (Honoré 1964: 470) A (particular) 'can' statement refers to a presumed actual state of affairs pertaining at the moment of utterance. Far from denying the antecedent, it requires the *possibility* that the antecedent be true, as Honoré acknowledges:

> [T]he statement that an agent can (particular) do so-and-so seems to require at least the truth of the statement that the agent can try to do so-and-so. (*ibid.*: 474)

The distinction between Honoré's can (general) and can (particular) is an Aspectual distinction as described in Section 2.1.5. Can (general) forms a verbal group with Iterative aspect, whereas can (particular) forms a verbal group with Dynamic aspect.

Metaphysical free will as described in Section 2.2.1 requires the conjunction of the following two propositions:

5. *A can* bring it about or ensure that *E* occurs.

AND

6. *A can* bring it about or ensure that *E* does not occur.

in each of which the modal verb *can* has a *realis* meaning, has Dynamic aspect, and expresses the modality type = 'Ability'. As noted in Section 2.2.3, *can* in the present tense is always *realis*, and Honoré's can (particular) also has Dynamic aspect. But the *can* in Honoré's analysis does not involve the modality type = 'Ability'.

After noting that an agent's (particular) ability to do so-and-so requires that the agent can try to do so-and-so, Honoré analyses the modality of the *can* in "can try". He claims, in essence, that for it to be true that an agent "can try" to do *E* it must be true that he can intend to do *E* and there must be some intermediate action *Y* such that doing *Y* is a means for doing *E*, and the agent can do *Y*. He then reasons as follows:

> What sort of 'can' is involved in 'can try?'. If 'can' is particular, we may, as we press the inquiry back, come fairly quickly to an initiating action such as the agent could not do, because he was prevented from doing it. Alternatively, if the initiating action is interpreted instead as an intermediate *Y* action, we may pursue the inquiry further into the past. Very often, however, we shall ultimately be faced with a question involving 'prevent' … (*ibid.*: 474).

Thus, it seems on Honoré's analysis, that if a proposition of the form "*A* can (particular) do *E*" is to be falsified, it will be falsified by some external circumstance which either prevents the doing of *E* or prevents the doing of some intermediate action for achieving *E*. If not falsified, a proposition of the form "*A* can (particular) do *E* involves an infinite regress and is strictly speaking never verified unless *A* actually does *E*. In practice, affirmative judgements are justified by treating one of the *cans* as can (general), and relying on evidence from similar cases.

On such an analysis, it becomes apparent that the modality expressed by the original *can* is not of type = 'Ability', but of type = 'Possibility'. It is governed, not by capacities inherent in the agent, but by the absence of preventing circumstances outside the agent. Alternatively, if we take the can (particular) to be governed by a

can (general), then the can (particular) is evaluated by our knowledge of what has happened on numerous similar occasions, and takes on an epistemic character .

Freedom characterised by Honoré's particular "can" may well be compatible with determinism or causal closure, but it is not the concept of freedom described in Section 2.2.1. Honoré himself says (*op. cit.*: 478) that his explication of "can" is compatible with determinism. In a case where he chooses beer over whiskey, he claims he could have chosen whiskey so long as he was not conscious of any factors that would have prevented him from choosing whiskey. I assume that "choosing" at least includes a physical act, like reaching out for the glass. Then if that act were in fact determined by physical events and laws beyond Honoré's consciousness, and had Honoré, before reaching out, asserted a conjunction such as "I can (particular) choose beer over whiskey and I can (particular) choose whiskey over beer", he would necessarily have spoken falsely, since one of the conjuncts would be false. It matters not that he would not know which until after the event.

3.7.5 Kratzer on 'Must' and 'Can'

Kratzer (1977) provides a useful insight into the semantic properties of the modal verb *can*. Although principally concerned with the modal verb *must*, she offers a semantic analysis of both *can* and *must* in terms of possible worlds. She begins with examples of 'must' used to express duty (deontic must), knowledge (epistemic must), dispositions (dispositional must) and strong desire (preferential must). She notes a range of variants within each of these broad categories, and from that concludes that "[a]ll this leaves us with many different 'must's' *and 'can's'*". (Kratzer 1977: 339, emphasis added) The extrapolation from 'must's' to 'can's' is not explicitly justified, but seems to rely implicitly on the association of these words with necessity and possibility, and the relation between these two logical concepts.

After examining various sentences employing 'must' for deontic, epistemic, dispositional and preferential purposes, Kratzer suggests that all such uses are implicitly qualified by a phrase commencing with "in view of …". Thus deontic 'must' means something like "must in view of one's obligations", epistemic 'must'

means something like "must in view of what is known", and so forth. She then asserts:

> "Similar considerations hold for the word 'can'". … "[R]elative modal phrases like 'must in view of' and 'can in view of' should be considered as the foundation of the modals must and can respectively." (*ibid*.: 342).

Kratzer's paper is not directed to the question of compatibilism, but in one example (*ibid*.: 343), she considers the proper interpretation of "could have acted otherwise" in the context of a fictional murder trial. If the question of whether the defendant could have acted otherwise is answered in view of "…the whole situation of the crime, which includes of course all the dispositions of the murderer (*sic*.)", the judge might be forced to answer in the negative. But that, according to Kratzer, would be a misunderstanding. The right understanding of such a question is "Given such and such aspects of the situation, could the murderer have acted otherwise than he eventually did?"

I agree with Kratzer that the CAN appearing in questions of legal culpability is likely to carry an implicit 'in view of …'qualification, and the same is true of many other uses of CAN. Kratzer's implicit 'in view of …' qualifications are reminiscent of Moore's suggestion that CAN should be understood conditionally. Wherever CAN is relative to an 'in view of …' qualification, it should be possible to recite limited sets of circumstances that allow a statement of possibility to come out true.

But Kratzer does not show that *all* uses of CAN are relative to such 'in view of …' qualifications. In particular, she does not show that a CAN expressing Dynamic modality of type = 'Ability' is relative in that way. Dynamic modality is grounded in the actual world. If an 'in view of …' qualification were to be inserted, it would have to be something like "in view of the totality of all circumstances within the backward light cone of the agent at the moment of utterance".

3.7.6 R.E. Hobart's Supercompatibilism

Another defender of compatibilism wrote under the name of R.E. Hobart. In a frequently quoted paper, Hobart argues that free will is not only compatible with determinism, but "inconceivable without it". Hobart says that to "tamper" with the meaning of either phrase "would be unpardonable", but it is clear that the concept of "free will" that concerns Hobart, though a natural and important concept, is a very different concept from the metaphysical free will described in Section 2.2.

Like Moore, Hobart is concerned with free will as required for "responsibility, merit and demerit, guilt and desert [as conceived] in life and in law and in ethics". Such free will, he says, implies that "… after an act has been performed, … one 'could have done otherwise'" (Hobart 1934: 1). As previously noted, the locution "could have done otherwise" is ambiguous. Outside metaphysical debate it normally expresses an *irrealis* kind of modality, by referring to the agent's ability in circumstances whose actuality the speaker does not assert. Less commonly, as when discussing metaphysical free will in Section 2.2 , "could have done otherwise" expresses a *realis* modality, having regard to the agent's capacities in the actual circumstances that obtain when a choice is made.

In describing an act of choice, Hobart reflects on his capacity as an agent, at the moment of choice:

> We say , "I can will this or I can will that, whichever I choose". Two courses of action present themselves to my mind. I think of their consequences, I look on this picture and on that, one of them commends itself more than the other, and I will an act to bring that about. I knew that I could choose either. That means I had the power to choose either. (Hobart 1934: 8)

Here, by focusing on the agent's perspective approaching an act of choice, Hobart seems to be describing something rather like what I call metaphysical free will. Metaphysical free will requires a *realis* modality of type ='Ability' with dynamic aspect. Hobart seems to be describing a *realis* modality with dynamic aspect, and the word "power" is usually more suggestive of Ability than externally-governed Possibility . But Hobart elaborates on his notion of power as follows:

> What is the meaning of "power"? A person has a power if it is a fact that whenever he sets himself in the appropriate manner to produce a certain event that event will actually follow. …I have the power to will so and so; that is, if I want, that act of will will take place. That and none other is the meaning of power.
> … Thus power depends upon, or rather consists in, a law. …Wherever there is a power there is a law. In it the power wholly consists. A man's power to will as he wishes is simply the law that his will follows his wish.

This elaboration shows that Hobart's concept of a power is not something internal to the agent, and the modality elucidated through that concept is not of type ='Ability'. Power merely provides a causal link between an agent's preferences and actions, and does not explain the source of the preferences. If the agent's preferences are to "up to" the agent, we would need something other than Hobart's concept of power to paraphrase the agent's freedom to form those preferences. But Hobart seems satisfied to treat preferences as something beyond the agent's control. Summing up, he asks:

> What … does … freedom mean? It means the absence of any interference with .. this [connection between preference and action]. Nothing steps in to prevent my exercising my power.
> All turns on the meaning of "can". "I can will either this or that" means, I am so constituted that if I definitively incline to this, the appropriate act of will will take place, and if I definitively incline to that, the appropriate act of will will take place.

The determinism concerning Hobart is not expressed as a causal thesis about events or states of affairs, but as the determination of a moral agent's actions by his character. Praise and blame are descriptions of an agent's character, and are justified in respect of acts only in so far as those acts are determined by that character. So analysed, free will and determinism are indeed closely related, but the concept of freedom he finds sufficient to justify praise and blame is very different from the concept of metaphysical free will described in Section 2.2.1.

3.7.7 Schlick's Compatibilism

Moritz Schlick (1962) is perhaps best understood as a contextual compatibilist. He treats free will as a problem in ethics, but for him, ethics is itself a descriptive science, which "seeks nothing but knowledge". Its aim is to explain those aspects of human behaviour which relate to moral judgements, and as such, it is a branch of psychology. Schlick is a physicalist and a believer in causal closure of the physical domain. He writes in 1930[27] as an avowed determinist, although he admits he cannot prove the principle of causality:

> [E]very explanation of human behaviour must … assume the validity of causal laws. … All of our experience strengthens us in the belief that this presupposition is realized. … Whether, indeed the principle of causality holds universally, whether , that is, *determinism* is true, we do not know; no one knows. (*loc. cit.*: 144)

Schlick introduces his brief chapter on responsibility and free will with expressions of "hesitation and reluctance", since he regards the free will problem as a pseudo-problem. Within his context of ethical enquiry, the only form of freedom he needs is freedom from compulsion. He claims that those who see freedom as incompatible with causal determinism are confusing two kinds of laws, prescriptive laws and laws of nature. Compulsion is a concept to be found only in prescriptive laws. Causal laws, in contrast are descriptive. Schlick equates an agent's responsibility with society's readiness to reward or punish the agent. He maintains that society's practice in rewarding and punishing presupposes causality, since we would not attribute responsibility for an act which arose purely as a matter of chance.

Here, he expressly assumes that there is no middle ground between a causally determined event and a chance event:

---

[27] And thus before much of the empirical evidence for indeterminism was available, although Popper (1982, p. 42) was arguing for indeterminism as early as 1925.

> [C]hance is identical with the absence of a cause; there is no other opposite of causality (*ibid.*: 156)

That assertion is similar to the assumption made by Smart, discussed in Section 1.3.2. It is an assumption that libertarians would dispute, and it seems that Schlick can appeal to societal practices as evidence for causality, only by first assuming its truth. I would argue that societal notions of responsibility equally assume that the agent to be rewarded or punished "could have acted otherwise" in a genuine sense, and not just in Schlick's hypothetical sense that he would have acted otherwise in counterfactual circumstances.

Schlick's notion of freedom as the absence of compulsion may be sufficient to explain society's practices of reward and punishment, but is far removed from the metaphysical conception of freedom described in Section 2.2. If we were to express an agent's freedom from compulsion in terms of CAN, the CAN would express the modality type ='Possibility' or *nihil obstat.* I differ from Schlick, not in denying that his *nihil obstat* freedom is compatible with determinism, but in his refusal to consider[28] that agents have a stronger "metaphysical" freedom, whose possession is necessary to account for our feeling that some acts are up to us.

Far from ignoring the subjective standpoint, Schlick claims what he calls "the consciousness of responsibility", to be "a welcome confirmation" of his position "that the subjective feeling of responsibility coincides with the objective judgement". Subjectively, an agent recognises not only that he took the steps involved in performance of an act but is aware "that he did it 'independently' [or] 'of his own initiative'".

> This feeling is simply the consciousness of freedom, which is merely the knowledge of having acted of one's own desires. And "one's own desires" are those which have their origin in the regularity of one's character in the given situation, and are not imposed by an external power …. The absence of the external power expresses itself in the

---

[28] As a founder of logical positivism, Schlick eschews metaphysics. His quarrel is with other philosophers who treat free will as an ethical problem. He "… exclude[s] only Bergson … with whom [free will is] a metaphysical problem [and whose] ideas … will not stand epistemological analysis [and] are of no significance for us".

> well-known feeling … *that one could also have acted otherwise*. (Schlick 1962: 154).

The *could* appearing in the last line does not express a *realis*, simple past tense, but an *irrealis*, Hypothetical modality, as Schlick makes clear in the next passage:

> It is of course obvious that I should have acted differently had I *willed* something else; but the feeling never says that I could also have willed something else, even though this is true, if, that is, other motives had been present. And it says even less that under *exactly the same* inner and outer conditions I could also have willed something else. … The feeling is not the consciousness of the absence of a cause, but of something altogether different, namely, of *freedom*, which consists in the fact that I can act as I desire.

The *should* in the first line of the above passage is of the modality type =□'Hypothesis' (Figure 2.1), as is the *could* of line 2, in the circumstances Schlick allows as true. The *could* of line 5 expresses a modality more like that involved in my conception of metaphysical free will. It seems to refer to (and deny) a power of the agent on a particular past occasion, in the circumstances which actually obtained. It is therefore the simple past tense of CAN, having *realis* meaning, Dynamic aspect, and expressing the modality type =□'Ability'. But Schlick denies that agents have such an ability, and is not concerned with metaphysical freedom. The aspect of *can* in the final sentence is uncertain, but it is also *realis* and expresses the modality type =□'Ability'. While I agree that (on some occasions with both Dynamic and Iterative aspect) "I can act as I desire", I reject Schlick's assertion that freedom *consists* in that fact.

3.7.8 Dennett's Compatibilism

Unlike most compatibilists, Daniel Dennett denies that free will is properly analysed in terms of 'could have done otherwise' (1984, 1984a), but he recognises the centrality of 'can' to the free will debate, and offers a useful analysis of various modalities expressed by that auxiliary (1984: 147 ff.).

Dennett is an outspoken compatibilist, but his expressed concern is with the moral and social issues that concern people in their dealings with the world. In two major works (1984 and 2003) and numerous papers he discusses concepts of freedom,

but none of the concepts of freedom he considers important are in any sense metaphysical.

Dennett is dismissive of metaphysical questions about freedom. They are not what interests him. In his recent book *Freedom Evolves,* (Dennett 2003: 296) he briefly refers to the argument that "… in a deterministic world at time *t*, nothing *can do* anything other than the one thing it was determined at *t* to do", but contrasts that conclusion with what he calls the "obvious fact that people today *can do* more than people used to be able to do." Stressing the "…explosive growth of *can-do* in recent human history" and the importance of "… *this* kind of 'can'" in contemporary ethical debate, he concludes that:

> The sense of "can" that has the moral import is not the sense of "can" (if there is one) that depends on indeterminism.

Undoubtedly, Dennett has described two very different senses of CAN. The CAN in his "… obvious fact that people today *can do* more than people used to be able to do." is very different from the CAN of metaphysical freedom. Like all CANs in the present tense it is *realis*, but its aspect is Iterative, referring to the collective abilities of all of humanity over recent history, and not than the ability of a particular agent at a particular time. Its modality type on the Possibility-Ability scale is unclear, but many of its instances would be of type ='Possibility'. And merely showing that a non-metaphysical sense of CAN has moral import does not justify the conclusion that no other sense of CAN is relevant to moral questions, implied by the first two uses of the definite article in the passage quoted above.

In the same book, Dennett expresses a "naturalistic" view of the relationship between philosophy and empirical science, proclaiming:

> … that philosophical investigations are not superior to, or prior to, investigations in the natural sciences, but in partnership with those truth-seeking enterprises, and that the proper job for philosophers is to clarify and unify the often warring perspectives into a single vision of the universe. (*ibid.*: 15)

These are sentiments of which I approve[29]. The account of free will that Dennett develops throughout several of his works is a naturalistic one, but I think he fails to place it within a "single vision of the universe".

Dennett's argument for compatibilism relies on his concept of "Intentional systems", as described in a (1971) paper by the same name. An Intentional system is one whose behaviour can in principle be explained and predicted in terms of mental states such as desires and beliefs. Paradigmatic examples of Intentional systems are people, some kinds of animals, and in some applications, computers.

Dennett's insight is that there are three different "stances" we can adopt when trying to explain or predict the behaviour of an Intentional system, and he describes these using the example of a chess-playing computer. He first mentions the "design stance". A technician knowing how the computer is constructed and programmed could (again in principle) predict the computer's next move from its current state, without any consideration of the game of chess being simulated. Such a prediction would be made from the design stance. Secondly, Dennett mentions the "physical stance". To adopt the physical stance in relation to a computer would involve consideration of it as a purely physical system of metal, silicon and plastic in a complex arrangement, governed only by the laws of nature. Adopting the physical stance in relation to a computer's chess playing output would be impractical, even for a 1971 computer, but the physical stance may be appropriate in cases of malfunction: a blown fuse, for example. Thirdly, Dennett mentions the "intentional stance". We adopt the intentional stance in relation to an Intentional system when we explain or predict its behaviour by attributing desires and beliefs to it. The intentional stance is the one that would normally be adopted by someone playing against the computer: "If I move my rook, it will want to protect its king", and it can lead to successful predictions where the design and physical stances would be practically impossible.

As the computer example shows, the intentional stance can be usefully adopted for systems to which even those who take human consciousness seriously would not ascribe "genuine" beliefs and desires. By definition only Intentional

---

[29] I describe myself as a naturalist, but by no means a physicalist. I endorse the truth-seeking partnership between philosophy and the natural sciences, but I claim that the natural sciences will need to encompass an ontology which goes beyond that presently recognised by mainstream physics, if they are to account for free will.

systems can be explained or predicted from the intentional stance. The design stance is applicable to a wider range of systems, including but not limited to machines designed to perform certain functions. I successfully predict that when I touch the key marked @, an @ symbol will appear on the screen before me. In making that prediction, I did not need to consider the computer as an intentional system. (I hope I have not hurt its feelings by saying that.) Dennett points out that the design stance can also be applied to natural objects: "Heavy pruning will stimulate denser foliage and stronger limbs" is his example (1971: 88). The physical stance is applicable in principle to any system, Although complexity rules it out for predicting certain types of behaviour in designed or Intentional systems, the physical stance would suffice to predict the fall of a man or a bicycle off a cliff. (We may need the intentional stance to predict what the man would shout at us on the way down.)

Dennett argues that free will is compatible with determinism because the concept of free will is meaningful only at the design and intentional levels, whereas questions of determinism are confined to the physical level of reality. I think Dennett's distinction of the physical, design and intentional stances says something important about how people explain different kinds of events and relations but the stances we adopt for different evaluations, or even the stances available to us in respect of different systems, tell us nothing about the reality of those systems. Thus, I do not believe Dennett's compatibilism achieves his (2003: 15) goal of "… unify[ing] … warring perspectives into a single vision of the universe" .

A "single vision of the universe" can accommodate multiple *levels of explanation*, but such a single vision, in my view, cannot ignore contradictions or incompatibilities across those levels. Dennett maintains his compatibilism only by confining himself to *separate* visions of the universe at different levels of explanation. If he were to stand back and combine those visions into a single, *universal* vision, the incompatibility would re-emerge. An argument such as van Inwagen's (1975, 1983) Consequence Argument (Section 3.3) involves events and entities described at multiple levels, and in my view, it succeeds in showing that determinism (described at the physical level) is incompatible with free will described at the intentional level.

A concept arising at the intentional level, which Dennett says is often confused with determinism, is inevitability. According to Dennett, those who assume determinism implies inevitability "… get it flat wrong. … It doesn't", (2003: 25). I

agree with Dennett that ordinary talk of inevitability is not inconsistent with an assumption of metaphysical determinism. But that is because our ordinary talk of inevitability is relative to a model which (inevitably) is incomplete. As human agents, we are never aware of all circumstances which might intervene. We succeed in life because we sufficiently often take into account the circumstances which are likely to be relevant, but we are not infallible. Suppose I drop a glass on to a concrete floor from four feet above and it breaks. Once it had left my hand, one might say it was inevitable that it would break, but another might say it was not inevitable, since I could have caught it with my other hand, *or* there might have been an elastic net two feet above the floor, *or* any of an indefinite number of other circumstances might have obtained. It is only relative to *all* circumstances that we could strictly say the breaking was inevitable, and since we cannot enumerate all possible circumstances which might have intervened but didn't, we can never say strictly and without qualification that an event was inevitable.

I claim, however, that the gap between determinism and inevitability is an epistemic one. Suppose that the universe were deterministic. An imagined all-seeing observer with full knowledge of the laws of nature would then be able to tell us after the glass leaves my hand whether its breaking is inevitable, by taking account of all potentially relevant circumstances. Only if the universe is somewhat indeterministic could the breaking fail to be inevitable from the all-seeing observer's perspective. If, as suggested by quantum mechanics, the distribution of electron density in the glass molecules is probabilistic, there is a small but non-zero probability that as the shock wave propagates through the glass upon hitting the floor, the molecular bonds between pairs of atoms which under normal circumstances would be forced apart by the shock wave will just happen to be momentarily strong enough to prevent shattering. Or less improbably – and not deniably without begging the question against metaphysical free will – I might not have made up my mind yet about whether to catch the glass with my other hand.

3.8 Contextual Accounts of Compatibilism

Most arguments for compatibilism assume causal closure of the physical domain[30], and then try to explain the concept of free will in a way which is compatible. In this section, I consider two more subtle forms of compatibilism which respect the simple man's concept of free will, but resist incompatibility by providing a contextual analysis of one or other of the seemingly incompatible concepts. Hawthorne (2001) offers a contextual analysis of free will, while Menzies (2003) offers a contextual analysis of causation

3.8.1 Hawthorne's Freedom in Context

Hawthorne models his contextual analysis of causation on Lewis's contextual analysis of knowledge claims. In his essay *Elusive Knowledge*, Lewis claims to steer a course "between the rock of fallibilism and the whirlpool of scepticism", by what he calls a "*sotto voce* proviso":

> *S knows* that *P* iff *S*'s evidence eliminates every possibility in which not-*P* – Psst! – except for those possibilities that we are properly ignoring (Lewis 1996: 551-2)

Just which circumstances may be properly ignored depends on the context in which the truth claim is evaluated. We may know we have hands when we go shopping for gloves, but not when we are engaging in epistemology[31]. Lewis emphasises that the context for what may be properly ignored depends on the evaluator of the truth claim, not on its assertor.

---

[30] Standard arguments assume physical determinism, which entails causal closure of the physical domain.

[31] Although Lewis makes no reference to Wittgenstein, it is tempting to regard the variety of contexts as language games. In *On Certainty*, Wittgenstein (1979) examines knowledge claims within language games.

Lewis admits at the end of the paper that his "Psst!" proviso is pragmatically self-defeating. By referring to the possibilities in which not-*P,* he is failing to ignore them[32]. He suggests that this difficulty could be overcome by semantic ascent to a meta-language, but leaves it to the pedantic reader to work out the "tiresome" details:

> If you want to hear my story told that way, you probably know enough to do the job for yourself. If you can, then my informal presentation has been good enough. (Lewis 1996: 567)

Hawthorne adopts *sotto voce* provisos to steer between the common sense assertions such as '*S* did *x* freely" and what he calls the "God's eye premise" that the world is largely deterministic[33]. Here is his analysis:

> S does *x* freely only if S's action is free from causal explainers beyond S's control – Psst! – apart from those causal explainers that we are properly ignoring. (Hawthorne 2001: 68)

Hawthorne uses the notion of a "causal explainer" to mean "any state of affairs which provides an adequate causal explanation of an action". He does not tell us what is meant by "adequate", but it seems that any adequate causal explanation would need to specify a set of conditions which are together sufficient for the effect.

In spite of the syntactic similarity, it seems to me that Hawthorne's "Psst!" provisos and Lewis's "Psst!" provisos are radically different. In the knowledge case, we ignore a potentially infinite number of possibilities in which not-*P*, while having no reason to expect any one of them to be actual. But if we take Hawthorne's God's eye premise seriously, at least one of the causal explainers we choose to ignore would have to pertain in the actual world. In this case, semantic ascent will not provide an escape, since the ultimate subject of our enquiry is not itself semantic or mind

---

[32] I'm not convinced he needs to make this concession, although it does not affect the subject of my present paper. Surely the exception is for all possibilities in which not-*P,* and we can safely use universal quantifiers without committing ourselves to the existence claims.

[33] Hawthorne offers no argument in support of the "God's eye premise", apart from an appeal to authority (though not God's authority): "… no sooner do we engage in the metaphysics of freedom – the systematic philosophical examination of free will – then we meet a compelling argument that we are never, or rarely free" (Hawthorne 2001, p.67)

dependent like knowledge, but – if we are realists about causation – a mind-independent fact about the causation of physical events.

Hawthorne's concept of acting freely is unlike most of the accounts considered. It does not expressly refer to what the agent *S can* or *could* do, but a notion of the agent's ability is contained in the reference to "causal explainers *beyond* S*'s control*". If *S* does *x* freely, then the only "causal explainers" we are allowed to consider must have been within the agent's control. But if there is a "causal explainer" within *S*'s control, and if *x* is determined, it seems that the effect of *S*'s control could not have been other than *x*. So whatever the nature and extent of the freedom inherent in the concept of *S*'s control, it is not the metaphysical freedom of Section 2.2, whereby *S* must be equally able to do *x*, and to ensure the non-occurrence of *x*.

3.8.2 Menzies' Causation in Context

While standard versions of compatibilism and Hawthorne's contextual version try to avoid incompatibility by redefining or restricting the context of free action, Menzies does so by offering a contextual account of causation. He proposes that account as a response to Jaegwon Kim's exclusion argument against non-reductive mental causation (Kim 1999). According to the exclusion argument, if mental states supervene on physical states without being identical to them, and if the physical domain is causally closed, then mental states as such cannot cause physical effects. Menzies points out that the exclusion argument relies on a tacit assumption (the "exclusion assumption") that, except in rare cases of overdetermination, no event has more than one complete causal history. According to Menzies, the exclusion assumption is false, since the concept of causation itself is context-dependent.

Causal judgments are invariably made relative to a model or context, in which some factors are attended to while others are ignored. Across all contexts, Menzies claims there is a unified *concept* of causation in terms of counterfactuals, but when that concept is applied in different contexts, different possible worlds count as the closest to the actual world, allowing a plurality of mutually independent causal pathways to be identified.

> The account … allows that a single event—in our example, the event of an individual's raising an arm—can have two different complete causal histories. Relative to the intentional model, the agent's reasons may be truly said to cause the agent's raising of the arm. Relative to the neurophysiological model, it is neurons' firing in the motor cortex which may truly be said to cause the behaviour. These are different causal relations, consisting in distinct processes, which can coexist because they are posited by two models, neither of which excludes the other. (Menzies 2003:. 219-20)

Menzies does not address the question of free will as such. He assumes, for the purposes of exposition, that the (neuro)physical domain is causally closed, and on his analysis, any physical action will have a complete physical cause. He makes no assumption about the closure or otherwise of the psychological domain, leaving open the possibility that since actions are allowed to have multiple causal histories, perhaps free choice can be accommodated within psychological models, while neurophysical models remain deterministic.

For anyone who is convinced of the causal closure of the physical domain, or is merely unwilling to consider that the physical domain is not causally closed, the Menzies account of mental causation should be attractive. Not only does it offer a way of attributing causal efficacy to mental states, but it might also accommodate the intuition that some actions result from free choice. Not being constrained myself by a belief in causal closure of the physical domain I do not need to rely on the Menzies account to validate the intuition of free choice. Unfortunately for those so constrained, I suspect a development of the Menzies account will not show free action to be compatible with causal closure since any determinism at the physical level would need to be reflected at the psychological level.

As a piece of epistemology, Menzies' account is attractive. When people describe and interpret events in the world, they make causal statements. The entities and relations referred to in causal statements depend on the matters known to the observer, and on the observer's pragmatic interests. Causal explanations are bound to be incomplete – many *conditiones sine quibus non* are unknown to the speaker, while many more are taken for granted. But causal explanations often succeed – in the sense of satisfying a meditator or an enquirer – and equally valid alternative explanations can often be given by assuming and attending to different subsets of

events and states of affairs thought to be relevant. In other words, causal *explanations* are undeniably contextual.

Menzies' account goes beyond causal explanations, however. He offers a unified *concept* of causation, applicable at all levels, and not just at the basic level of physics. In that respect, he expressly distinguishes his account from Jackson and Pettit's view on program versus process explanations (Jackson & Pettit 1990):

> They express their view in terms of causal explanation rather than causation, but that difference is not crucial here. They distinguish explanations of physical effects in terms of mental states —program explanations — from proper causal explanations of physical effects in terms of earlier physical effects — process explanations.(Menzies 2003: 201)

Menzies is highly critical of philosophers who claim that causation exists only at the level of fundamental physics:

> I take this to be *a reductio ad absurdum* of the [exclusion] argument. It flies directly in the face of commonsense and scientific thought to say that the special sciences do not investigate and discover real causal structures. This view makes a mockery of the enormous efforts devoted in the special sciences to formulating experimental and observational methodologies for testing causal hypotheses. It would follow from this position that all these efforts are misdirected because they could not, by definition, reveal anything about the nature of causal processes. There must be something wrong with the argument if it leads to this highly implausible result. (*ibid.*: 199)

I respectfully disagree. Experimental chemistry, for example, proposes and tests causal hypotheses expressed in terms of entities such as acids, bases, benzene rings, and valencies. The chemist draws a molecular formula representing some yet-to-be-synthesised acid, say, and predicts that it will react with some alcohol to produce a previously unknown ester likely to have properties – density, volatility, fragrance – of a certain kind. Yet if pressed, as a typical reductionist[34], he would concede that the chemical reaction involves redistributions of electronic charge according to some physical laws. These laws, he might say, could in principle be described in terms of Schrödinger wave functions, though as a chemist he would not attempt to do the

mathematics himself. The physicist down the corridor might tell us that the wave functions only disclose probabilities. If we dig deeply enough, we might find there is no strict causation going on at the microphysical level, but the probabilistic laws enable him to tell us what we will observe when a sufficient number of similar atomic systems are treated in the same way. Certainly, the chemist with his test tubes of liquids can proceed *as if* there is causation at the level of molecules, and can develop his experimental and observational methodologies to deal with specimens of liquids, solids and gases. He can give causal *explanations* at that level, without having to believe in autonomous chemical causation.

On the Menzies account, there are multiple levels of causation, none of which is any more fundamental than any other. Menzies denies "…that the concept of causation is the concept of a categorical, absolute relation". Rather:

> we conceptualise [causal relations] as entities occupying certain functional roles that are defined with respect to abstract models. Recognition of this *fact* opens up the possibility of seeing that there can *be* different levels of causation. There may be a level at which mental states cause behaviour by way of distinctive psychological pathways; and a different level at which physical brain states cause behaviour by way of distinctive neural pathways. *These different levels of causation need not be in competition with each other*. (Menzies 2003: 196, italics added)

If we treat Menzies' account as epistemological, the last sentence can be taken to mean only that we *need* not attend simultaneously to different levels. But unless humans are somehow constituted as to be *unable* to consider events and processes across different levels, the only way the levels can fail to compete with each other is by failing to entail any inconsistent facts within any domain in which comparisons can legitimately be made. For a "global realist" such as myself (Section 1.1.4), one such domain would be the entire actual world.

Even if causal processes at the neurophysical and psychological levels do not compete with each other, they are intimately related by the effects they are each supposed to produce. Presented with an effect *E*, we may be able to describe non-competing causal pathways at the neurophysical and psychological levels, converging

---

[34] I suspect most chemists of my generation are reductionists, at least about chemistry.

on that effect. And such descriptions at the psychological level can invoke such notions as the agent having made such and such a choice. But if we entertain the notion of metaphysical free will as characterised in Section 2.2, the psychological level allows the agent a choice between alternatives. At some stage before *E* does or does not occur, the agent *S* is in a psychological state □ in which it is both true that *S* can cause *E* and true that *S* can refrain from causing *E*. Each *can* in the previous sentence expresses a *realis* modality of type ='Ability' with dynamic aspect. It refers to an actual capacity of the agent, at the time of choice, in respect of a particular possible event. Put simply, at the psychological level, whether *E* occurs is "up to" the agent. In contrast, if the neurophysical level is causally closed, the probability of *E* is fixed. If on one occasion the two independent levels of causation converge on either the occurrence or non-occurrence of *E*, that would be an unremarkable co-incidence. If the two independent levels of causation converge to the same result whenever an agent believes he is making a choice, that would be miraculous.

Alternatively, should it be argued that the physical facts, and hence the probabilities, would be different at each successive occasion, we could consider a large set of possible worlds with identical histories up to the moment the agent makes a choice. In any one world, the correspondence of the actual outcome with whichever choice the agent makes would be consistent with that outcome being probabilistically determined, but it would be remarkable if the actual outcome in *every* possible world happened to correspond with the counterpart agent's choice in that world.

3.9 Conclusion

In Section 3.6, I have argued that metaphysical free will as defined in Section 2.2 is incompatible with causal closure of the physical domain. If I am right, then one cannot consistently maintain both that human agents (or for that matter, anything in the universe) have metaphysical free will, and that the physical domain is causally closed. Faced with such an inconsistency, many would deny that metaphysical free will exists. As argued in Section 2.3, I believe there are good reasons to recognise metaphysical will, and unless there should turn out to be even better reasons to

believe in physical causal closure, I would deny that the physical domain is causally closed.

In Section 3.5, I considered views representative of many philosophers who have described senses of freedom that do seem to be compatible with a deterministic universe, and that are probably no less compatible with a universe whose physical domain is causally closed. Many have assumed, and some such as Dennett have expressly argued, that such senses of freedom are the only senses we need, and that even if metaphysical free will is a real phenomenon, it can safely can be ignored. In the last chapter of his book *Elbow Room: The Varieties of Free Will Worth Wanting*, Dennett (1984) asks "Why do we want free will?" and suggests that the only reason human agents want free will is to "secure for us our dignity and responsibility" (*ibid.*: 153). These concepts of dignity and responsibility are meaningful only at the intentional level, so if any kind of freedom were to be postulated at the physical level, it would not be of a kind worth wanting.

Dennett is entitled to his opinion on what is worth wanting, but what is worth wanting does not determine what exists. If people do not *have* "free will" in some sense in which they desire it, the question of *why* people want free will in that sense may be worth exploring. But if people *do* have "free will" in the relevant sense, the question of why they want what they have has no bearing on the fact they have it. If I want angels' wings (which I don't have), a psychoanalyst might fruitfully obtain a research grant to examine hypotheses about what prompts that desire in me, but such research would have no bearing on whether or not I can fly. If I want toes (which I do have), my desire is equally irrelevant to whether I have them. That hypothesis can be tested by removing my boots – or to my own satisfaction by wiggling them.

My desire to have what I *do* have can be explained best in terms of my aversion to counterfactual circumstances in which I am deprived of that property. Consistently with humans having free will, their *desire* to have free will might be explained by showing that the absence of free will would be undesirable. And indeed, that is just what we find in so many discussions by compatibilists. If people didn't have free will, societal practices of praise and blame, punishment and reward, and the concept of desert would have no basis. But our wanting such practices to have a basis cannot affect the ontological question of whether free will in the desired sense exists.

From my perspective, whether people have metaphysical free will is an interesting question about the world, principally because the fact of its possession would provide us with information about the causal structure of the universe: namely that not all events have their probabilities determined by prior physical states and events in combination with physical laws. But for those whose philosophical interest is motivated primarily by moral questions such as the justification of praise and blame, the existence of metaphysical free will should also be an important issue, if the very processes of deliberation in which we engage when deciding what we should do (or judging what an agent should have done), presuppose that, at the end of the deliberating process, we shall have (or the agent had) the ability to give effect to the conclusion of the deliberations.

While few would deny that people act as if they have free will, some, such as Wegner (2002) maintain that the feeling of acting under the idea of freedom in the metaphysical sense, is no more than an illusion. Unless there is some independent reason to believe metaphysical freedom is impossible, such as a proof that the physical domain is causally closed, I submit that the "best" explanation for the feeling of freedom is that agents in fact have a capacity to exercise real choices. In chapter 4, I shall argue that physics itself gives no reason to believe that the physical domain is causally closed.

CHAPTER 4

CAUSAL CLOSURE OF THE PHYSICAL DOMAIN

4.1 Introduction

In Chapter 3, I argued that metaphysical free will in the sense described in Section 2.2.1 is not only incompatible with physical determinism, but incompatible with the causal closure of the physical domain. The compatibilism debate is usually framed in terms of "free will" and determinism, sometimes with a footnote conceding the falsity of determinism but claiming the author's position would be sustainable in more words if that "complication" were allowed for. The real antithesis of free will or free action is the physical causal closure principle, which does not assume determinism, but would be a corollary of determinism if the world were physically deterministic.

Though few philosophers nowadays would insist that the physical domain is deterministic, many if not most contemporary philosophers addressing the question of free will within the analytical tradition accept that the physical domain is causally closed in a sense to be clarified below. But a belief in causal closure of the physical domain seems to come at a high price. Either one must abandon the folk notion that we can sometimes exercise choices that cause physical events, or one must find a way of reconciling the capacity for free choice by human agents with all events having their chances determined by purely physical antecedents.

The arguments presented in Chapters 2 and 3 show that neither of those prices is acceptable. My purpose in the present chapter is to discover why a belief in causal closure of the physical domain is so popular. I begin with an attempt to trace the

sources of that belief in twentieth century philosophical literature, and to identify any affirmative arguments for it. I then consider a number of arguments against the possibility that the physical domain is open to non-physical influences. I conclude that physics itself provides no reason to believe that the physical domain is causally closed. Therefore in the light of good evidence and sound arguments in favour of metaphysical free will and of the demonstrable incompatibility of metaphysical free will with causal closure of the physical domain, causal closure of the physical domain should be rejected.

4.2 Causal Closure of the Physical Domain

4.2.1 The Causal Closure Principle

Causal closure can be expressed in terms of whatever relata figure in one's favourite account of causation. Menzies (2003: 197) expresses the principle in terms of physical states. Baker (1993) defines it in terms of property instantiation. For the purpose at hand – an attempt to account for certain physical *events* – I take causal closure of the physical domain to mean that all "physical" events, unless they are uncaused, have at least their probability of occurrence determined by other physical events and states of affairs. This definition would allow an event to have a plurality of independent sets of causes so long as one of those sets were wholly physical, but for reasons given in Section 3.6.2, I do not believe that the existence of separately describable mental causes would overcome the incompatibility of metaphysical free will with a co-existing set of completely physical causes.

4.2.2 Hempel's Dilemma

If the causal closure principle is to be made precise, there must be a clear notion of what is meant by "physical". That issue has been discussed in the philosophical literature of physicalism, a doctrine that in most forms entails what I call causal closure of the physical domain.

Carl Hempel (1980) commented on the obscurity of the physicalist claim that the language of physics can serve as a unitary language of science without some specification of exactly what is meant by "physics". Depending on how the word "physics" is understood, physicalism can be made either trivially true or demonstrably false. The scope of "physical" in my description of causal closure can likewise be cashed out in various ways. On the broadest interpretation, "physical" can be taken to encompass whatever entities, properties and relations are needed to explain any occurrence. The proposition then becomes tautologous. At the other extreme, if "physical" is limited to the entities, properties and relations presently recognised by physics, the proposition is almost certainly false (Montero 1999, 2003).

Crane and Mellor (1990) argue on similar considerations that physicalism is incoherent. They claim that if "physical" refers only to the subject matter of pure physics, physicalism would obviously be false, but that there is no principled way of drawing a line between what are generally conceded to be the physical sciences, and the "human sciences" exemplified by psychology which deal with intentional states and emotions. If "physical" is extended to encompass the human sciences, physicalism becomes vacuously true. Thus, they contend, no serious question of physicalism arises.

Papineau (1993, 2001, 2002) defends a form of physicalism that he prefers to call "philosophical naturalism", and which entails causal closure of the physical domain. Indeed, his defence of physicalism expressly takes the "completeness of physics" or causal closure of the physical domain as one of its premises. To make his position explicit, Papineau (1993: 29-30) defines "physical" to encompass all of the entities, properties and relations needed to explain paradigmatically and uncontroversially physical events such as falling rocks, and he simply stipulates that it shall contain nothing which is fundamentally mental. A similar approach is taken by Montero (2003) in a paper discussing varieties of causal closure, where she avoids the ambiguity in "physical" by discussing the causal closure of "the fundamentally non-mental".

Excluding anything fundamentally mental from what is meant by "physical" leaves open the question of whether there is in fact anything fundamentally mental, and it probably captures the intuitions shared by those who argue about physicalism and causal closure. I would still prefer to characterise what is physical in an

affirmative way, leaving open the question of whether the physical domain encompasses everything there is, and more pertinently everything that contributes to the causation of physical effects. For my purposes, the physical domain needs to include all those entities and properties to which the theories either accepted or seriously discussed within current day physics refer. It also needs to include entities of kinds not yet discovered or postulated, so long as such entities share the same kinds of properties as are attributed to the entities currently known or postulated.

I suggest the characteristic feature of all physical entities is that they have mass and/or energy in some form or other. Current physical theories refer to matter and energy manifested as fields, particles, waves, strings or "branes" (see Section 5.12.8). Physical descriptions of matter and energy also attribute various properties such as spatio-temporal location, momentum, charge and spin to elements of matter and energy, but I think it is safe to say that any physical theory will ultimately deal with entities having mass and/or energy in some manifestation or other, and will seek to identify regularities in the observable behaviour of entities or systems of entities having mass and/or energy in terms of properties attributable to such entities.

Between them, the general theory of relativity (see Section 5.12.4) and the Standard Model of particle physics (see Section 5.12.5) seem to cover the gamut of physics as currently practised. But if some new particle or force-transmitting field were to be discovered, it would be natural to consider it part of the physical domain, so long as mass or energy were attributable to it. Likewise, if some novel property of a particle or energy field were found to play an explanatory role in physical theories, it would be likely included within physics. But if some entity were discovered or postulated that did not have mass or energy, it and any properties peculiar to it would be quite foreign to physical theories as we know them, and I would say such an entity would lie outside the physical domain.

Thus, I shall take the entities of the physical domain to be those having mass and/or energy. The language of physics also includes spatio-temporal concepts and properties other than quantities of mass and energy – such as charge or spin – but the bearers of such properties and the entities to which spatiotemporal location are attributable by physics will themselves have mass and/or energy. As I use the term "mental", mental events lie outside the physical domain, and my demarcation therefore seems consistent with those of Papineau and Montero.

4.3 Denying Free Will

If free action in the metaphysical sense and causal closure of the physical domain are indeed incompatible, one cannot consistently believe in both of these hypotheses. Faced with such a choice, a consistent incompatibilist must either deny the reality of the impression that he engages in metaphysical free will or give up his belief in physical causal closure. Many choose physical causal closure and jettison the reality of free action.

Merely by choosing one or other way, the incompatibilist seems to engage in a free action. Choosing to deny that one is capable of choosing would be a pragmatic refutation of that denial. Still, an incompatibilist who comes to believe firmly in physical causal closure might claim that he is not *really choosing* that belief over the alternative belief in free choice, but is somehow passively witnessing a sequence of changes in the state of his brain, caused by physical events beyond his control. That to me seems a difficult belief to maintain. I have experienced occasions when my thought processes were momentarily compelled by external factors, as when suddenly confronted with imminent danger. Those occasions felt vastly different from occasions of quiet philosophical reflection, where the subject and intensity of my attention from moment to moment does not seem to be forced on me at all.

The subjective difference between controlling one's though processes and having them controlled may, I concede, be part of a grand illusion. It is nevertheless central to the way in which I engage with the world, and I am reluctant to deny the reality of my choices unless the arguments and evidence in favour of causal closure of the physical domain are compelling. In this chapter I deny that there are any compelling arguments in favour of that doctrine.

4.4 The Burden of Proof

In Section 3.1, I accepted the burden of showing that metaphysical free will is incompatible with causal closure of the physical domain, and I claim to have

discharged that burden in Section 3.3. In Section 2.3, I offered two arguments to support the claim that human agents genuinely have a capacity for free will. As exemplified by Roget's Thesaurus, the structure of languages shows that humans at least share a basic *belief* that our species has a capacity for exercising choices that affect the world around us, but I have to allow that, as Dennett (1984, 2003) and Wegner (2002) would have it, such a belief may be a systematic and universal delusion. Independently of language, strong evidence for the reality of metaphysical free will comes from the pragmatically-undeniable fact that humans have consciousness, the evolution of which would be inexplicable if it did not have some efficacy in the physical domain.

Taken together, the reality of metaphysical free will and its incompatibility with the physical causal closure thesis show that thesis to be false. Yet many respected philosophers, accepting the incompatibility of free choice with some view of physics that entails causal closure, reject (or observe themselves coming to reject) the reality of free choice embodied in my concept of metaphysical free will. It is therefore incumbent on an advocate of genuine free choice to consider whatever evidence and arguments have been propounded in favour of causal closure. If the arguments or evidence for causal closure turn out to be stronger than the evidence and arguments I have given in Chapters 2 and 3, I would be forced to concede there is something wrong with those arguments, or with the overall inference from "*p* and *q* entails not *r*" and "*r*" to "either not *p* or not *q*".

4.5 Who are the Proponents of Causal Closure?

4.5.1 Not, as a Rule, the Physicists

My research suggests those who argue for – or merely assert – causal closure of the physical domain are primarily philosophers, or perhaps scientists working in the special sciences, several levels removed from fundamental physics. But even if no physicist has ever argued that the physical domain is causally closed, the absence of such arguments cannot be taken as evidence that physicists do not believe in that

proposal. It may simply indicate that active physicists do not choose to address the question.

There are plenty of reasons why physicists as a rule may not enter the debate on causal closure. For a start, there are plenty of theoretical and practical questions to interest them within physics. As Schwartz and Begley observe:

> Engineers who design or use transistors, which exploit quantum phenomena, rarely think about the ontological implications of quantum mechanics and whether the mind shapes reality; neither do high-energy physicists, as they work out the chain of reactions in a particle accelerator. For every hundred scientists who use quantum mechanics, applying the standard equations like recipes, probably no more than one ponders the philosophy of it. They don't have to. You can do perfectly good physics if you just "shut up and calculate," as the physicist Max Tegmark puts it. (Schwartz and Begley 2002: 296)

Secondly, professional physicists get paid to do physics. They survive and thrive by teaching and publishing physics. The causal closure question is not physics, but metaphysics. Respectable physical questions have answers that can be proposed and tested intersubjectively (see Section 6.5.4) The only empirical evidence any of us is likely to have for non-physical causes of physical effects is an awareness of our own capacity as agents to exercise a choice, but that evidence is of no use against the materialist who flatly asserts that our “awareness” is no more than an illusion, or that what we mean by “free” is not what we think we mean. Why would a physicist with the intellectual skills, training, laboratory resources and funding to address unsolved but potentially soluble physical questions spend his or her time and effort arguing with philosophers? Most prefer to leave the task to Lockean under-labourers.

4.5.2 The Assumption of Causal Closure

On the rare occasions when a credible physicist enters the debate, his or her views deserve to be taken seriously. But I have yet to find a physicist arguing in favour of causal closure. Much practical physics is conducted on the *assumption* that the system under investigation is causally closed. Indeed, practical physics tries, so far as possible, to work with systems that are shielded from extraneous causal influences,

and especially from interventions by human agents. In setting up a laboratory experiment to measure the behaviour of billiard balls, we would certainly keep the meddlesome child away from the table. Physicists *assume* causal closure of the systems they investigate for good practical reasons, but to make such an assumption is far from subscribing to a *belief* that the system is in fact closed. Still less does it imply a belief that causal closure of the physical domain applies universally.

As physics developed in the twentieth century, it came to be recognised that any physical system about which there can be knowledge cannot be totally isolated from human beings. At some stage, a human must interact with the system as an observer, and that at least raises the possibility of causal or other influences passing not only to the human as observer, but from the human as agent. Indeed many accounts of quantum mechanical events go further and actually require the involvement of an observer to "collapse the wave packet".

4.6 Theories Implying Causal Closure

Arguments expressly in favour of physical causal closure are hard to find. However, causal closure of the physical domain is entailed by or implicit in some popular theories about the causal structure of reality, variously labelled as "mechanism", "materialism" or "physicalism". The latter two terms in particular are used interchangeably by some, or given overlapping scopes by the same or different authors. For my purposes, it is not necessary to delimit any of these terms precisely, since they all have, at their core, a commitment to some kind of physical causal closure principle. It is that common core that I intend to challenge.

4.6.1 Mechanism

The term "mechanism" is rarely seen in the recent literature, although some of the positions defended and criticised under that name in the earlier literature are still defended as physicalism or materialism. More importantly, some of the arguments

raised against mechanism still stand as challenges to those newer doctrines. Historically, mechanism was the view or hypothesis that a full description of reality could be given in terms of mechanical interactions between material entities.

In its original form, mechanism was falsified by the recognition of electromagnetic properties and forces, which were not mechanical in the traditional sense. Once these non-mechanical properties and forces were recognised by physics, "mechanism" gave way to either physicalism or materialism.

4.6.2 Materialism

Materialism can be taken as a proposition that all facts about the world, and in particular all facts customarily described in mental language, supervene on facts about matter.

4.6.3 Physicalism

If physicalism is to be given a broader scope than materialism, it might be taken as a proposition that all facts about the world, and in particular all facts customarily described in mental language, supervene on facts about whatever entities and relations are required to describe physical facts. Papineau defends "physicalism" in his (1993) book *Philosophical Naturalism* and "materialism" in his (2002) boo*k Thinking about Consciousness*, but it is not clear that he intends to distinguish the terms as I have proposed, if at all.

4.7 Arguments Invoking Physical Causal Closure

4.7.1 Broad's Analysis

In the third chapter of *The Mind and its Place in Nature,* C.D. Broad (1925) defends what he calls "two-sided interaction" between mind and body. Unless the mind simply *is* the body or some part of it, two-sided interaction requires the denial of causal closure of the physical domain. Broad discusses various arguments against two-sided interaction, which involve or amount to arguments in favour of physical causal closure. He divides those arguments into philosophical and scientific arguments.

4.7.1.1 Philosophical Arguments

Broad refers to two philosophical arguments against interaction, namely:

> that minds and mental states are so extremely unlike bodies and bodily states that it is inconceivable that the two should be causally connected (*ibid*.: 97)

> that, whenever we admit the existence of a causal relation between two events, these two events (to put it crudely) must also form parts of a single substantial whole. (*ibid*.: 99)

Broad rejects each of these arguments in turn. He takes the fact of a "close correlation between certain bodily events and certain mental events" to be undeniable, but concedes "that concomitant variation is not an adequate criterion for causal connexion". Something further is required for causation, but Broad denies that either of the additional criteria presupposed by the above arguments is necessary. The requirement for a "high degree of likeness" between cause and effect is vague, and there are uncontentious causal connections in the physical domain (such as between

"draughts and colds") where the respective cause and effect or cause and effect are "extremely unlike each other". He claims the requirement that cause and effect be part of a single whole is not proven, and that in any case, it is possible that mind and body do form part of a substantial whole, being neither wholly mental nor wholly physical.

4.7.1.2 Scientific Arguments:

Broad also discusses two arguments against interaction based on empirical science, identifying them as:

> (1) The Argument from Energy. (*ibid.*: 103)
>
> (2) The Argument from the Structure of the Nervous System. (*ibid.*:□109)

The "Argument from Energy" concerns the law of Conservation of Energy, and Broad discusses it at some length. Proponents of the argument as discussed tacitly assume "that, if a change in [one system] has anything to do with causing a change in [another system], energy must leave [the first system] and flow into the second". But, as Broad points out, that is not what Conservation of Energy requires. All that law requires is that *if* the first and second systems together form a closed system, and *if* some energy is lost from the first system, a corresponding amount of energy will be gained by the second. The law of Conservation of Energy is a limiting case of the First Law of Thermodynamics, discussed below in Section 4.11.

The "Argument from the Structure of the Nervous System" begins with the recognition that the body is capable of producing many reflex actions in response to external stimuli. The physical systems involved in these reflex actions are the same systems as those that produce supposedly deliberate actions, and do so without any intervention from the mind. There is no gap in the causal chain between stimulus and action, and therefore no scope for the mind to intervene in the supposedly deliberate actions. Broad rightly rejects this argument, saying it confuses a gap in explanation with a spatiotemporal gap. There may, in such actions, be a spatiotemporal sequence

of neurophysiological events, but that would not mean there is no explanatory gap. Explanations of deliberate actions invariably make reference to factors beyond the neurophysical systems involved, without which there would indeed be an explanatory gap. I would further reject the argument from the structure of the nervous system, in that reflex actions leave no scope for free will, and therefore cannot be relevantly analogous to deliberate actions, that do.

Broad's consideration of the arguments against interaction is interesting, particularly as it predates the rise of strong forms of physicalism and identity theory in the mid-twentieth century. I expect most of the physicalists and identity theorists would have been aware of Broad's position, but were evidently unpersuaded by it. In the following sections, I shall consider the views of some of those later philosophers.

4.8 Parsimony or Simplicity

Perhaps one of the greatest attractions of the physical causal closure principle is its parsimony. According to Ockham's razor, entities, or at least entity *types*, should not be multiplied beyond what is necessary to explain the phenomena under consideration. Denying causal closure of the physical domain clearly implies the existence of at least one type of non-physical entity, and in deference to William of Ockham, we therefore should not deny such causal closure *unnecessarily*.

There are, however, many precedents for admitting new entity types to theories where a plea of necessity, to some degree or other, can be sustained. In the practical sciences, physics has expanded its ontology to admit particles such as neutrinos and entities such as electromagnetic fields. In doing so, it respected Ockham's razor, inasmuch as the newly postulated entities were necessary to provide coherent theories, or to account for empirical observations. While admitting new entity types as necessary, physics nevertheless advances by reducing the number of entity types where possible. Before Rutherford, physics recognised as many types of fundamental particles (atoms) as there were elements. It then appeared that all such atoms were composed of only two fundamental particles, protons and electrons. The ontology of elementary particles was later expanded to include neutrons, necessitated by recalcitrant data about atomic weights. The "fundamental" particles of the early

twentieth century are now said to be composed of a small set of particles even more fundamental, but theoretical physicists are willing to postulate types not yet observed to fill out their theories. Mathematics, too, has been willing to admit new entity types as "necessary" to build up a theory: negative numbers, rational numbers, irrational numbers, complex numbers.

4.9 Physical Causal Closure and Mental Causation

The philosophical literature on mental causation in the second half of the twentieth century is voluminous. A recent survey of the major threads in that literature is provided by Kim (1999: 28), who observes that just about all mainstream theories share two assumptions: firstly that mental states strongly supervene on physical states, and secondly that "any causal chain that involves at least one physical event must lie wholly within the physical domain". Baker (1993: 79) and Menzies (2003: 197) agree that the causal closure assumption is widely held[35]. If any cogent arguments in favour of causal closure exist, one might hope to find them in the literature surveyed by Kim.

Kim traces the debate back to the brain state identity theories advocated by Feigl (1958) and Smart (1959). According to Kim, these authors:

> helped set the basic parameters and constraints that were to come – a set of broadly physicalist assumptions and aspirations that still guide and constrain our thinking today. [Although brain state theories now have few adherents], almost all the participants in the debate have stayed with physicalism. (Kim 1999: 2)

---

[35] Kim poses the following problem for an account of mental causation :

> [I]f mental properties are physically irreducible and remain outside the physical domain, then, given that the physical domain is causally closed, how can they exercise causal powers, or enjoy any kind of causal relevance, in the physical domain? (Kim 1999: 58–9)

I think the structure of this question is instructive. The use of "if" presents the proposition that "mental properties are physically irreducible and remain outside the physical domain" as a hypothesis, perhaps to be discarded if we cannot find a satisfactory answer to the question. In contrast, the proposition "that the physical domain is causally closed" is expressed as "given", and thus more firmly entrenched in any revision of beliefs which may be required.

Because central state identity theories recognise nothing that is non-physical, physical causal closure follows trivially. Advocates of such theories need provide no separate argument for physical causal closure as such, but if Kim is right about the history, there would be no good reason for the assumption of physical causal closure to survive the death of identity theories, without some such separate arguments. Perhaps the arguments given by the identity theorists can be adapted to support causal closure without mind-brain identity.

Feigl's identification of mental and physical states stems from a strong commitment to physical determinism, and draws encouragement from apparent successes in scientific reduction in the first half of the twentieth century. Feigl rejects non-physical entities for lack of any persuasive metaphysical arguments or scientific evidence in their favour, and applies Ockham's razor against accepting unnecessary entities (1958: 386). He rejects epiphenomenalism, which would leave "'dangling' nomological relations" (*ibid.*: 382) between physical events and inefficacious mental events.

Smart (1959: 142) argues against irreducible psychical events, "[m]ainly because of Occam's razor". He refers to Oppenheim and Putnam's (1958) paper, "Unity of Science as a Working Hypothesis", and finds it "frankly unbelievable" that the eventual reduction of everything to physics will fail to accommodate sensations, leaving "'nomological danglers' to use Feigl's expression". He believes his identity (brain-process) theory provides an adequate explanation of behaviour and what we perceive as sensations:

> If it be agreed that there are no cogent philosophical arguments which force us into accepting dualism, and if the brain process theory and dualism are equally consistent with the facts, then the principles of parsimony and simplicity seem to me to decide overwhelmingly in favour of the brain-process theory (*ibid.*: 156).

In a later paper, Smart (1963: 651) defines "Materialism" as " … the theory that there is nothing in the world over and above those entities which are postulated by physics (or, of course, those entities which will be postulated by future and more adequate physical theories)". He expressly denies "… that in the world there are nonphysical entities and nonphysical laws", and in particular, he denies "the doctrine of psychophysical dualism … [and] any theory of 'emergent properties'". (*ibid.*:. 652).

In this paper, Smart gives a further argument for materialism, partly based on evolutionary theory. If there are any non-physical properties or entities, he asks how such an entity could have "suddenly" arisen in the course of animal evolution:

> A change in a gene is a change in a complex molecule which causes a change in the biochemistry of the cell. … But what sort of chemical process could lead to the springing into existence of something nonphysical? No enzyme will catalyze the production of a spook! (Smart 1963:. 660).

Smart is unimpressed by a suggestion that the non-physical entity comes into existence as a by-product of the complexity of the system. This would treat the non-physical as an emergent property of the type he vehemently rejects at the beginning of his paper. In an understatement contrasting in style with his earlier "spook" reference, he concludes that at "the very least, we can vastly simplify our cosmological outlook if we can defend a materialistic philosophy of mind".

4.9.1 Oppenheim and Putnam's Working Hypothesis

Oppenheim and Putnam's (1958) paper, cited by Smart, provides an optimistic, but by their own admission incomplete, account of reductionism. The "working hypothesis" to which their title refers is that a unified science can be described in terms of six "reductive levels", each of which but the lowest can be "micro-reduced"[36] to the one immediately below it. Those six levels are specified, in terms of the entities they describe, as follows:

[36] A micro-reduction essentially involves the translation of theories at the higher level into theories at the lower level, with the entities of the respective theories at the two levels standing in part-whole relations.

6 Social groups

5 (Multicellular) living things

4 Cells

3 Molecules

2 Atoms

1 Elementary particles

Oppenheim and Putnam allow that their working hypothesis can never be justified on other than empirical grounds, but they argue that on the evidence available at the time, it is supported by its simplicity, and by the proven success of reductions between adjacent pairs of levels, at least up to level 5. They argue that simplicity supports their model against rivals, such as "psychism and neo-vitalism", which postulate entities or properties at level 5 which cannot be explained at level 4, and which, at the time of writing, did not leave themselves open to empirical testing.

Oppenheim and Putnam expressly acknowledge that their empirical hypothesis may need to be amended or even abandoned, if some future discoveries are made which it fails to accommodate:

> On the other hand, if the efforts at micro-reduction should seem to fail, we cannot preclude the introduction of theories postulating presently unknown relevant parts or presently unknown relevant attributes for some or all of the objects studied by science. Such theories are perfectly admissible, provided they have genuine explanatory value. [In such cases] a new working hypothesis of micro-reducibility, obtained by enlarging the list of attributes associated with the lowest level, might then be correct. However, if there are presently unknown attributes of a more radical kind (e.g. attributes which are relevant for explaining the behavior of living, but not of non-living things), then no such simple "repair" would seem possible. In this sense, Unity of Science is an alternative to the view that it will eventually be necessary to *bifurcate* the conceptual system of science, by the postulation of new entities or new attributes unrelated to those needed for the study of inanimate phenomena. (Oppenheim and Putnam 1958: 12-13)

In his 1968 book, *A Materialist Theory of the Mind,* Armstrong argues for materialism "from the supremacy of physics", finding it "incredible" to accept "that the whole world studied by science contains nothing but physical things operating according to

the laws of physics *with the exception of the mind*." (Armstrong 1968: 49, italics in the original):

> Until relatively recently, it did not seem scientifically likely that biology was completely reducible to chemistry, and chemistry completely reducible to physics. Under those circumstances, it was quite plausible to think that psychology dealt with quite new entities and quite new laws …. But if all sciences except psychology are, in theory, very complex particular cases of the fundamental science of physics, it seems very unlikely that psychology is an exception. It would follow that some materialist theory of the mind is the true one. (*ibid*.: 50).

Armstrong sees no need to accommodate irreducible "emergent" laws, but unlike Smart, he allows that such laws could be incorporated into a materialist metaphysics, should science discover any in the future (*ibid*.: 359).

Lewis (1966) bases his identity theory on two premises, firstly that experiences are defined by their causal roles and secondly that a true and exhaustive account of all physical phenomena can (in principle) be given by a unified body of scientific theories, ultimately expressible in terms of fundamental particles, fields and similar entities of kinds now known to physics.

The second premise, which Lewis describes as a "plausible hypothesis" is a commitment to the causal closure of the physical domain. It does not deny the possibility of non-physical phenomena, but if there are any such phenomena, they could not be experiences as defined in the first premise. Lewis seems to rest his confidence in the second hypothesis on the cumulative success of the physical sciences in explaining a multitude of phenomena. He expressly relies on the account of reduction given by Oppenheim and Putnam (1958).

As Kim notes, it was identity theories such as those of Smart and Feigl which introduced physicalist assumptions (including causal closure) into the mind body debate. As type-type identity succumbed to the multiple realisation argument, repairs were made within those assumptions, as "… even those who had a major hand in the demise of the Smart-Feigl materialism have continued their allegiance to a physicalist world view" (Kim 1999: 2). Causal closure of the physical domain, already made explicit by Lewis, could be accommodated within non-reductive theories, and mind-body supervenience took the place of strict identity.

Because these physicalist assumptions were uncontroversially shared, it is not surprising that subsequent authors within the paradigm felt it unnecessary to restate the arguments underlying them. Yet such justification as was originally given for causal closure by the identity theorists seems to have faded away with the identity theories themselves.

4.10 Inductive Arguments from Physics

The history traced by Kim explains why many philosophers unquestioningly embrace the principle of causal closure, but does not justify the principle itself. A minority of contemporary philosophers expressly address the question, and look to the success of physics within its own domain to justify the principle.

Melnyk (2003) argues as follows:

> [A]lthough the claim that the physical is causally closed is not explicitly stated in physics textbooks, it may nonetheless be inferred from claims that *are* explicitly stated in physics textbooks. According to the textbooks, then, contemporary physics has succeeded in finding sufficient physical causes for physical effects of very many kinds; and it has found no physical effects at all for which it is necessary (or even likely to turn out to be necessary) to invoke non-physical causes. (Melnyk 2003: 160-161)

Melnyk goes on to reject the suggestion that an inductive extrapolation from laboratory systems to brains is unwarranted, because he claims brains are merely physical structures "…which seem, from the physical point of view, to be quite unexceptional", and whose [physically realised] "basic biochemistry is apparently no different from that of the cells of other types". In opposition to Melnyk, I would argue that the inductive extrapolation is indeed unwarranted. However similar the underlying physics of brain and other tissues may be, it is hard to deny that brains perform or participate in functions like perception and conscious phenomenal experience not achieved by other lumps of meat. Perhaps those differences in function can be explained in purely physical terms, but if so, brains will have to be recognised as having properties radically different from the properties of a kilogram

of prime fillet steak. If perception and conscious phenomenal experience are no more than physical properties, there would need to be something fundamentally different about those few structures in the world that are capable of bearing those properties.

Pace Melnyk, there seems to be one class of physical effects for which physics as presently understood cannot offer an explanation, namely the immediate physical effects of conscious free choices made by human agents. Within the theories of physics, physical effects are either determined or have their probabilities determined by prior physical events and states of affairs, together with the laws of nature. To explain why an agent acts one way, rather than another, it is necessary to invoke a special science such as psychology. While physicalists may claim that a psychological explanation ultimately supervenes on some set of physical facts, such a claim has no empirical basis, but seems to *assume* causal closure. Physicists, in contrast to physicalists, do not pretend that physics can explain conscious free choices.

4.11 Non-Closure and the First Law of Thermodynamics

The first law of thermodynamics has been amply corroborated for all kinds of mechanical systems. According to one standard textbook (Young 1992: 488), it states that the change in the internal energy ΔU of a thermodynamic system is equal to the heat added to the system minus Q the work done by the system W, or equivalently

$$Q = \Delta U + W.$$

For an isolated system – one which does no work and exchanges no heat with its environment– the internal energy remains constant ($\Delta U = 0$). For a system which does no external work and has no work done upon it, the first law reduces to the conservation of energy.

When a thermodynamic system undergoes a change from one state to another, its internal energy, U, may change, but there is overwhelming empirical evidence that the energy difference, ΔU, between two states is independent of the path by which the change occurs. On free expansion of a gas at constant temperature, no heat is

absorbed and no work is done, but if the same gas is allowed to expand slowly against a weighted piston, heat needs to be absorbed from its environment to maintain temperature, and work is done. As well as these two alternatives there are a multitude of other possible paths between the same initial and final states, admitting free variation of Q and W, while maintaining the same ΔU.

When a human agent acts, a multitude of nested thermodynamic systems are involved. These range from the agent in combination with his environment[37] down to his brain, or whatever part of it is involved in the triggering step of the overt action. Whichever thermodynamic system we consider, there is no reason to suppose that the change in the internal energy of that system is not balanced by any heat absorbed and any work done.

As I shall be arguing elsewhere, determinism does not hold for token quantum events. Although quantum theory can accurately predict the statistical frequency of different possible outcomes where a large number of events of the same type occur, there is no known way of predicting the outcome of a single measurement. If an agent's free choice were to involve merely a selection between alternative quantum possibilities which then trigger a chain of macroscopic events, there would be no inconsistency with the first law, since any one of the possible outcomes is presumed consistent with that law.

The processes of free choice to which I allude would involve causation, but not the transfer of energy. Within purely physical systems, it is true that causation usually involves the transfer of energy or some other physical quantity, leading some philosophers to offer general accounts of causation in terms of transference of conservable quantities (Dowe 1995). But the systems I am considering are not purely physical systems. They attribute causal efficacy to mental entities that are non-physical, in the sense of not possessing mass or energy and not being capable of transferring or receiving mass and energy. Some specific theories in which non-physical minds are supposed to affect selection among possibilities allowed by quantum mechanics will be described in Chapter 6.

---

[37] Strictly, there is no boundary for the environment. If the agent raises his arm, as agents in the free will debate are wont to do, he changes the centre of mass for the Earth. If he does so outside on a starry night, he absorbs photons from distant galaxies.

Whether or not there are non-physical causes is a matter of controversy. But irrespective of whether there *are* non-physical causes, the very concept of a non-physical cause seems to preclude its occurrence contravening the first law. However "physical" might be defined, we can assume that anything having mass or energy is "physical". Thus anything involved in the transfer of heat or the performance of work is *not* non-physical. The reader may wish to deny for other reasons that non-physical things exist, but whether or not there are non-physical things, it should be agreed that no non-physical things transfer heat or do work.

4.12 Non-Closure and Conservation of Momentum

Conservation of energy is a corollary of the first law of thermodynamics where no work is performed. Papineau (2002: 236) claims that even if the scalar quantity of energy were conserved where non-physical causes operate, the vector quantity of momentum would not be conserved.

Papineau makes that point in a historical discussion contrasting Leibniz with Descartes, and admits that he is using the word "physical" in a restricted sense appropriate to the age. But although his own conception of "physical" is open-ended, his defence of physical causal closure assumes that any causal influence from a non-physical mind would involve some kind of *sui generis* force. Forces are vectors, and any net force on a physical system would alter its momentum. According to the way I characterised "physical" in Section 4.2, any entity that exerts a force would *ipso facto* be a physical entity. According to Papineau's own delineation, "physical" expressly excludes anything fundamentally mental, so the concept of a force-exerting non-physical "mind" whose efficacy he denies is one I would agree is not to be found.

Papineau's arguments for causal closure consistently assume that any causal influence on a physical system must involve a force[38]. By demonstrating the

---

[38] It is clear that Papineau intends "force" to be understood in a mechanical sense, rather than a metaphorical sense (like the "force" of an argument). In an endnote to his (2001) paper, he elaborates:

> [I]f you regard forces as otiose, [you may] think of the circumstances that "cause forces" as themselves directly causing the resulting accelerations.. In that case, you will replace the question of whether there are 'mental forces' with the question of

impossibility of "vital forces", "mental forces" or other kinds of non-physical forces, he assumes he has also demonstrated the impossibility of non-physical causal influences.

But there is no reason to assume that causal influences on a physical system must be *via* a force. The mental influences I postulate are non-physical in the sense that they do not have mass or energy, and are therefore incapable of exerting forces. Several models discussed in Chapter 6 postulate mind-brain interaction not involving forces.

Irrespective of the model proposed, if the mind causally affects the physical brain, without exerting any kind of force but merely by selecting among alternative states permitted by quantum theory, there would be no violation of the laws of conservation of momentum and conservation of energy. We can assume any physical brain event operates according to the conservation laws. The non-occurrence of such a brain event would equally conserve energy and momentum, and (unless the universe is strictly deterministic) we can equally envisage alternative brain events (or tokens of an identical brain event at different times), any one of which would conserve energy and momentum should it occur. All that free choice requires is the ability of an agent to choose among physically lawful alternative event tokens.

Other arguments that causation by a non-physical "mind" would contravene conservation laws (Mohrhoff 1999, Wilson 1999) appear to be directed against particular models and hypotheses. In this chapter, I do not defend any particular model but merely argue that non-physical causes as such are not precluded. Several interactionist models are discussed and compared in Chapter 6.

4.13 Non-Closure and the Second Law of Thermodynamics

I have argued that causal action of a non-physical mind on a physical brain need not contravene either the First Law of Thermodynamics (conservation of energy) or the Law of Conservation of Momentum. I now consider whether such interaction would

---

> whether specifically mental initial conditions … make a difference to accelerations. (2001: 34)

necessarily conflict with the Second Law of Thermodynamics, about which Sir Arthur Eddington has famously said:

> The … Second law of Thermodynamics – holds, I think, the supreme position among the laws of nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equation – then so much the worse for Maxwell's equation. If it is found to be contradicted by observation – well, these experimentalists do bungle things sometimes. But if your theory is found to be against the Second Law of Thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation. (Eddington 1933:□4)

The Second law of Thermodynamics still holds a major position in natural science. It can be expressed in various ways. A recent physics textbook offers the following as alternative statements of the Second Law:

> It is impossible for any system to undergo a process in which it absorbs heat from a reservoir at a single temperature and converts the heat completely into mechanical work, with the system ending in the same state in which it began (Young 1992: 514).
>
> It is impossible for any process to have as its sole result the transfer of heat from a cooler to a hotter body (*ibid.*: 515).

Other statements of the Second Law employ the concept of entropy:

> When all systems taking part in a process are included, the entropy either remains constant or increases. Or: No process is possible in which the total entropy decreases when all systems taking part in the process are included. (Young 1992: 526)

A practical handbook (Bosch 1993: 69) describes entropy as "a measure of the thermal energy in a system which is no longer capable of performing work". Entropy may be thought of as the degree of randomness or disorder in a system, and the law tells us that the entropy of any closed system must increase (or at least cannot decrease).

Entropy has been described by one critic of mind-body dualism, Morowitz (1987) as "that most hard-nosed steam engine concept", It appears in thermodynamic equations as a calculable physical quantity, having the dimension of energy divided by temperature. Although absolute entropy cannot be measured, entropy differences are quantifiable in physical systems such as engines and heat pumps. The concept of entropy was introduced in 1865 by Clausius, who defined it by the relationship:

$$dS = dQ/T,$$

where dS is the incremental change of entropy in a system to which heat of dQ is supplied in a reversible infinitesimal process at temperature T. A reversible process is an idealisation which can never be achieved. For any irreversible process, only the following inequality holds:

$$dS < dQ/T.$$

To approximate the entropy of a system, therefore, one would need to imagine the system first at a temperature arbitrarily close to absolute zero (0ºK), and supply heat at an infinitesimal rate until the temperature of T was achieved. The entropy would then be given by

$$dS = \int_0^T dQ/T.$$

According to Penrose[39] (2004 : 690), although Clausius introduced the notion of entropy, it was Bolzmann in 1877 who "… made the definition of entropy clear (or, at least as clear as it seems possible to make it)."

---

[39] Although Penrose takes a philosophically unorthodox position on mind-body interaction as discussed in Section 6.6, the book from which I quote in this section does not deal with the mind issue at all, but is a survey of contemporary physics for the lay reader. Interestingly, its title seems to have changed just before publication. A book review from *The Times* published in *The Australian* on 4th August, 2004 gives the subtitle as "A Complete Guide to the *Physical* Universe", and it is listed with that subtitle on the website of amazon.co.uk. The book as

On Penrose's account of Bolzmann's analysis, the entropy *S* for a physical system is given by:

$$S = k \log V,$$

where *k* is Bolzmann's constant (1.38×10^{-23} Joules per Kelvin), and *V* is a measure of the degrees of freedom in the system. While that equation looks simple, the concept of degrees of freedom is elusive.

Reichenbach (1971: 55) gives the Bolzmann equation as

$$S = k \log W,$$

and describes *W* as "the probability of a state". Intuitively, if we imagine a system such as an ideal gas in a sealed container, some states of that gas are more probable than others. It is much more probable that we would find the gas molecules to be randomly distributed and moving at random velocities, than that we would find them all in one corner of the container, or all moving in the same direction. States far from equilibrium are highly improbable, and therefore have a much lower value of *W* than states close to equilibrium. They therefore have lower entropy.

Penrose's account is more precise, and relies on a description in terms of state space. We are to imagine a physical system, such as a gas in a container, as fully described (classically) by a finite, though enormous, number of real number parameters. Thus, the instantaneous state of an ideal gas containing *N* molecules is fully described by 6*N* parameters, specifying the three spatial coordinates and three velocity (or momentum) coordinates of each of the molecules. Every possible state of the system can thus be treated as a "6*N*-tuple" of real numbers, defining a point in a 6*N*- dimensional space, and the dynamic evolution of the system can be represented by the trajectory of a point through that state space.

---

supplied by Amazon bears the subtitle "A Complete Guide to the *Laws of the* Universe". Has Penrose lately decided that the physical universe is all there is, or has he been censored by a physicalist publisher?

Because each parameter of each molecule is assumed to take a real number value, the number of points in the state space is infinite (with cardinality C). Each of those states would have zero probability, and no single state could be said to be more probable than any other. (A similar paradox arises when a pointer is spun and allowed to come to rest. If the angular displacement from some arbitrarily chosen direction can take any real value from 0 to 360º, the probability of its stopping at any position is zero, yet the total probability must sum to 1. For practical purposes, we divide the circle traced by the pointer into finite segments as small as we choose, and assign probabilities according to the size of those segments.) To be able to assign probabilities in the 6*N* state space, we must imagine that space to be divided into a finite number of "boxes" containing sets of points whose values can treated as equivalent to each other, or "macroscopically indistinguishable". There will be vastly more states in each of the boxes containing states close to equilibrium than in boxes containing highly ordered states. In that sense, a state in a box containing states close to equilibrium is more probable, and such a box is said to have a greater volume *V*, than a box containing fewer indistinguishable states. So understood, the volume *V* forms the argument in Penrose's version of Bolzmann's equation.

The division of the state space into boxes of "macroscopically indistinguishable" states is referred to as "coarse graining". Coarse graining involves an inescapably arbitrary choice as to what amount of latitude is allowed in treating the values of (classical) parameters such as position and energy of two particles as the same, and prevents any absolute quantification of entropy. In practical applications, absolute entropy is never measured, but only differences in entropy between two states of a system. Although the arbitrary nature of coarse graining prevents any precise quantification of entropy, Penrose shows, the concept is "robust" inasmuch as any "reasonable" coarse graining will give an entropy value of the same general order.

4.13.1 Ambit of the Second Law

Pace Eddington in the passage quoted at the beginning of this section, I claim that a perceived conflict between a theory of mental causation and the Second Law need not be fatal. In the first place, the law applies to macroscopic, closed systems. Secondly,

it is open to question whether it properly applies to living systems at all, let alone those systems in which conscious choice seems to play a causal role.

Like any proposed natural law, the Second Law of Thermodynamics must be expressed in declarative language, and any purported statement of it has a truth value. Injudiciously expressed statements of any purported law may be falsified, while more cautious and qualified statements may survive and be taken as true[40]. If the ambit of the Second Law should properly be restricted to macroscopic, closed inanimate systems, it would be silent about mind-body interaction, and certainly would not preclude it. But some insights of information theory, matching quantified information to negative entropy, prompt speculation that the Second Law might be extended to describe systems encompassing both mental and physical subsystems, whereby an increase in informational entropy compensates for a decrease in physical entropy. Such speculations are discussed in Sections 4.13.6 and 4.13.7.

4.13.2 Entropy and Life

Eddington may be right that a purported "theory of the universe" generally disagreeing with the Second Law would be hopelessly flawed, but mental events occur in only a minute part of the universe. They occur, or at least have their most immediate effects, in living organisms, which are by no means closed systems. A theory of mind is not a "theory of the universe" to echo Eddington's phrase, and it might turn out that the Second Law does not apply to mental processes, or perhaps even to living systems in general.

Although textbook statements of the Second Law refer to (physical) systems without restriction, and although the bodies of human agents are no doubt physical systems, it seems that all empirical evidence for the Second Law comes from the

---

[40] Empirical evidence in the Southern hemisphere might lead to the postulation of a "law" that (*ceteris paribus*) water going down a plughole will spiral in a clockwise direction. It's certainly more than an accidental generality, but the statement is falsified by the behaviour of suitably regular systems north of the equator. The fault lies with the statement of the law, which can be revised either by limiting its domain to the Southern hemisphere, or by relativising the direction of rotation to that component of the Earth's rotation along an axis parallel to that of the drain.

study of physical systems in which minds play no agentive role. Indeed, Brillouin plausibly claims that there can be no evidential justification for applying the second law to any system containing living organisms – conscious or otherwise:

> How can we compute or even evaluate the entropy of a living being? In order to compute the entropy of a system, it is necessary to be able to create or to destroy it in a reversible way. We can think of no reversible process by which a living organism can be created or killed: both birth and death are irreversible processes. (Brillouin 1949: 564)

One of the earliest statements of what is now recognised as the Second Law comes from Lord Kelvin (William Thomson):

> It is impossible, *by means of inanimate material agency*, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects (Thomson 1852b, italics added)

The words in italics are of at least historical significance. Kelvin clearly did not regard the Second Law as describing (or governing) systems involving an animate agent. In a passage attributed to him by Ehrenberg[41] (1967), he is more explicit:

> The animal body does not act as a thermodynamic engine. The means in the animal body by which mechanical effects are produced cannot be arrived at without more experiment and observation. …Whatever the nature of these means, consciousness teaches every individual that they are, to some extent, subject to the direction of his will. It appears therefore that animated creatures have the power of immediately applying to certain moving particles of matter within their bodies, forces by which the motions of these particles are directed to produce derived mechanical effects.

In contrast to his exclusion of animate agents in his statement of the Second Law, Kelvin recognised that living creatures were subject to the First Law:

---

[41] I have been unable to find this passage in any of Kelvin's papers. An Internet search shows the passage quoted by several other authors citing Ehrenberg, but Ehrenberg gives no source, and with him the trail goes cold.

> The question, "Can animated creatures set matter in motion in virtue of an inherent power of producing a mechanical effect?" must be answered in the negative, according to the well-established theory of animal heat and motion, which ascribes them to the chemical action … experienced by the food. (Thomson 1852a: 507)

The mere fact a "father of the Second Law" limited its domain to non-living systems does not mean he was right to do so. Many sons and daughters have made their mark in domains unimagined by their fathers, and perhaps the Second Law has a universality beyond Kelvin's expectation.

It is true that statements of the Second Law as found in physics textbooks do not exclude systems involving mental agency. That omission is not surprising in a textbook on physics, however. Physics deals with physical systems, and not every statement in a physics textbook needs to remind us that it is a statement about physical systems. If the world contains entities that are non-physical, the laws of physics would not purport to describe their behaviour.[42] Not all physicists are physicalists, and the physical domain gives them plenty to investigate, without metaphysical speculation about mind-body relations. Kelvin, though renowned as a physicist, was a polymath by today's standards. His published works extend over a wide range of the physical, biological and even social sciences. Perhaps he speaks to us from a more leisurely age, in which a working scientist could keep abreast of broader issues, and concern himself with the forest, as well as the trees. Nowadays, the exponential growth of knowledge forces physicists to specialise[43], and only a few such as Stapp (Section 6.5) and Penrose (Section 6.6) choose to specialise in the peripheral question of how the mind relates to the physical world.

---

[42] My analysis rejects the assumption that mental states are nothing more than physical states, but reliance on such an assumption to defeat an argument against causal closure would be circular.

[43] In 2002, the shelf space required for the printed version of just one journal, *Physical Review*, grew by nearly five metres!

4.13.3 Statistical Nature of the Second Law

Quite apart from any special attributes of mental systems, the world abounds with systems whose entropy can be seen to decrease, and such decreases can readily be explained without doubting the universality of the Second Law. For a start, and as recognised in Bolzmann's derivation, the Second Law is a statistical law. For any dynamical system, there is a finite and in principle calculable probability that it will spontaneously change to a state of lower entropy, over some period of time. As Maxwell observed in 1870[44], the second law has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you can't get the same tumblerful of water out again.

4.13.4 Maxwell's Demon and The Fluctuation Theorem

In his 1870 letter to Lord Raleigh, quoted above, James Clerk Maxwell acknowledged the statistical nature of the Second Law. The following year (1871: 308) he published a thought experiment featuring what Kelvin later dubbed Maxwell's "demon", to emphasise the same point:

> [T]he second law … is undoubtedly true as long as we can deal with bodies only in mass, and have no power of perceiving or handling the separate molecules of which they are made up. But if we conceive a being whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are still as essentially finite as our own, would be able to do what is presently impossible to us. For we have seen that molecules in a vessel full of air at uniform temperature are moving with velocities by no means uniform, though the mean velocity of any great number of them, arbitrarily selected, is almost exactly uniform. Now let us suppose that such a vessel is divided into two portions, A and B, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only

---

[44] In a letter to J.W. Strutt (Lord Raleigh), quoted in Leff and Rex (1990, p. 290)

> the swifter molecules to pass from A to B, and only the slower molecules to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics.

The possibility of a Maxwell's demon would appear to contradict those statements of the second law which talk of "impossibility". But we don't need to call on the demon to falsify such statements. According to statistical mechanics, there is an extremely small but non-zero probability that a temperature gradient will occur spontaneously between A and B. The point was made by Poincaré in 1893[45]:

> .. to see heat pass from a cold body to a warm one, it will not be necessary to have the acute vision, the intelligence, and the dexterity of Maxwell's demon; it will suffice to have a little patience.

One hundred years later, Evans *et. al.* (1993) calculated the probability of observing a decrease of entropy in small physical systems over measurable time intervals[46]. Their "Fluctuation Theorem" was experimentally verified in 2002, for trapped particles of a few microns diameter, for periods of the order of a second (Wang *et. al.* 2002).

4.13.5 Open Systems

One doesn't need either the patience of Poincaré or the sophisticated apparatus of Wang *et. al.* to observe systems in which entropy decreases. When I retrieve from the fridge the can of beer I put there yesterday, its entropy has decreased. Grains of sand washed upon a beach form orderly patterns. Birds build nests, and people spend large parts of their lives, for better or worse, creating order from disorder. None of this is surprising. Nor is it contrary to any reasonable statement of the Second Law, because the systems in which the entropy is seen to decrease are not isolated systems. We are happy to assume that the decrease of entropy within the observed system is more than

---

45 "Mechanism and Experience", quoted in Leff and Rex (*loc.cit.*, p. 291)

46 The authors refer to "violations" of the Second Law. Strictly speaking, they have "violated" only an expression of the law which fails to acknowledge its statistical nature.

offset by an increase in entropy in some larger system of which the observed system is a subsystem.

As evidenced by the sorting of sand grains on a beach, life is not essential for localised and sustained entropy decreases to occur, but within the realm of everyday experience, living organisms provide the most striking examples. Schrödinger (1967: 67 ff.) describes the "characteristic feature of life" as its ability to maintain itself by "extracting 'order' from the environment", or "feed[ing] upon negative entropy". All such processes, he stresses, take place within the laws of physics, including the Second Law, but the processes he describes are metabolic, and non-mental. It is not until an Epilogue (*ibid.*: 86 ff.) that Schrödinger discusses the apparent interaction of mind and body, and the apparent contradiction between his physical account of life and his "incontrovertible direct experience" that he can control some of his bodily motions. Schrödinger's response is a form of panpsychism, giving priority to a unified and universal consciousness, on which the physical world depends. While I am not persuaded to accept his ontology, I note that he does not claim that the action of mind on matter is governed by the laws of physics.

As emphasised by Schrödinger, living organisms are not closed systems. For as long as life persists, they reduce their own entropy by discharging entropy into their environment. To observe the Second Law in action, we must consider a larger system, including the organism and many elements of its environment, including but not limited to the low-entropy food it consumes, the work it performs, the thermal energy it radiates and the higher entropy matter it excretes. By including enough of the environment, we may arrive at a practically closed system, in which the Second Law is observed. But the living body alone is not a closed system, and whatever the relationship between mind and body, we can be equally sure that mind and body do not form a closed system without at least as much of the environment as is required to supplement the body. Thus, there is no more reason to deny that the entropy of a mind-body subsystem can decrease than to deny that the entropy of a living body decreases.

In Section 6.3, I shall discuss the thirty-year project of Nobel prize-winning brain physiologist, Sir John of Eccles, to give an account of free will consistent with empirical science. At an early stage of that project, Eccles collaborated with Popper, and one question which concerned them was the consistency of mind-body interaction

with the First Law of Thermodynamics. But Popper and Eccles saw no similar difficulty with the Second Law. In one of the dialogues mentioned in Section 6.3.3, Popper observes:

> I do not think that we have to worry about the Second Law of Thermodynamics at all. We have only to assume that the brain gets tired under mental activity and that this tiredness is in some way or other equivalent to heat production, and so to a degradation of energy, and that the Second Law is thus preserved. There is just a lot of heat produced by all these processes: one gets, so to speak, a hot brain. (Popper and Eccles 1977: 541).

Statements of the Second Law like those quoted above from Young are splendid examples of scientific hypotheses, in a form of which Popper would surely approve. Boldly they make a negative existential claim, challenging the doubter to falsify them. Although these statements of the law do not exclude the intervention of immaterial minds, I think it is safe to say that the impressive body of empirical data providing corroboration for the Second Law is drawn from experiments on closed physical systems, shielded from external influences. I do not for one moment suggest that spoon-bending or other psychokinetic tricks provide any credible falsification of the Second Law as broadly stated, but only that in so far as these statements purport to extend beyond the purely physical domain, they are unsupported. If one assumes that the physical domain is causally closed, then *a fortiori,* a non-physical mind can not cause a process whose sole result is the transfer of heat from a cooler to a hotter body, But if the universal statement of the Second Law is derived from such an assumption, that statement cannot be used to support a belief in causal closure.

4.13.6 Information and Entropy

While I claim that the Second Law gives no grounds for denying that a non-physical mind could causally affect a living physical system, some twentieth century work on information theory invites speculation that in an expanded form, the Second Law might even be applicable to mind-body interaction.

When the Second Law is stated in terms of physical entropy, it is hard to see how it could be a property of a non-physical mind. Physical entropy has dimensions of energy divided by temperature. Both energy and temperature are themselves physical quantities, and are not attributable (other than metaphorically) to a mind, which has neither mass nor spatial location. The fact that we cannot envisage a non-physical mind contributing or absorbing entropy to or from a physical system seems of itself another reason to say that mind-body interaction cannot be precluded by the Second law.

But in a broader sense, entropy has been understood as a measure of randomness or disorder. Mental processes involve understanding or imposing order on abstract thought and ideas, and if there are aspects of mind which are not wholly underwritten by physical states such as brainstates – as I would claim – then degrees of order can be properties of non-physical minds.

While the Clausius definition of entropy is hard to reconcile with mental contents, Bolzmann's statistical definition is more amenable. An ordered state of a physical system is less probable than a disordered state, because there are numerically more ways in which the system can be disordered. Analogously, a piece of intelligible information resembles an ordered state of a physical system, because it excludes a vast number of inconsistent possibilities.

The link between information and entropy was noted in 1929 by Leo Szillard. Szillard described a thought experiment in which a Maxwell's demon was able to decrease the entropy of a system, but only by a process of acquiring information about the state of that system (Szillard 1964).

4.13.7 Maxwell's Demon and Information

Developing Szillard's idea, Brillouin (1952) argues that Maxwell's demon is impossible in principle. To operate the trapdoor, the demon must acquire information about the trajectories and velocities of the individual molecules. If the demon literally "sees" the molecules, it must absorb photons from them at a minimum energy above the background radiation of the chamber, and if it remains in thermal equilibrium with

the gas, its own entropy must increase by more that the decrease of the entropy of the gas.

Morowitz builds on Brillouin's ideas to conclude that "Mind body dualism is in direct contradiction to the second law of thermodynamics" (Morowitz 1987: 275). He contends that any non-material mind which operated to select between energetically equivalent alternative microstates of a physical system would, in effect, be a Maxwell's demon, since it would need somehow to acquire information about those microstates. Morowitz claims that, however it is acquired, information about a system is intimately related to entropy:

> The Brillouin analysis was based on information theory and theory of measurement. … [T]he development of information theory led to the notion that even "entropy", that most hard-nosed steam engine concept, was at its root mentalistic bearing on the observer's knowledge of the microstate of the system (*ibid.*: 273)

Although Morowitz contends that "mind body dualism is in direct contradiction to the second law of thermodynamics", that does not lead him to reject mind body dualism. Rather, he suggests that the Second Law might need to be limited to situations not involving mind:

> I don't think that this would radically effect (*sic*.) physics, but it makes a large difference to biophysics. *It allows us to think of mind as related to the kind of knowledge a system has of its own state without taking a measurement*. (*ibid.*:□275, italics added)

Although Morowitz does not cite Kelvin's words from 1852, the two seem in agreement about the inapplicability of the second law to mental causation. And so long as the Second Law is stated in terms of physical entropy, I would be happy to agree with them both. I am nonetheless prepared to speculate on whether the Second Law might be expanded to admit a wider concept of entropy, and in a way compatible with mind-body interaction.

Rodd (1963) relates the statistical concepts of entropy and information, and shows a mathematical similarity between entropy and an observer's lack of information about the state of a thermodynamic system. He argues with reference to an expanding gas system that entropy change and information acquisition are

"physically related", and derives the following generalised version of the Second Law of Thermodynamics:

$$\Box(S - U) \Box 0,$$

where S is a measure of the entropy of a system, and U is a measure of an observer's uncertainty about the state of the system.

Attempts by Rodd and others to link information quantitatively with entropy have been widely criticised. The strongest objection seems to be that entropy (or at least entropy change) is supposed to be an objective property of a dynamic system, whereas information is inescapably a subjective property of a particular observer. That seems to me to be a telling argument for systems accessible to independent observers, but if interactions take place between a subject's mind and body, it is conceivable that in such interactions, a reduction in the entropy of the physical brain is offset by a reduction in the uncertainty in the agent's mind. I admit this idea is highly speculative, and my main claim, that the Second Law does not preclude mind-body interaction, does not depend on it.

4.14 Empirical Arguments

In their introduction to Part II of the recent collection "Physicalism and Mental Causation", Walter and Heckman (2003: 133) observe that "…a violation of the causal closure of the physical should [in principle] be detectable and confirmable on empirical grounds, since if the physical is *not* causally closed, the occurrence of some physical events would have to be explained by reference to non-physical causes and the laws governing them". On its face that looks like a strong argument against causal closure, although I doubt the authors intend it as such. I claim there is an important class of physical events that cannot be explained *except* by reference to non-physical causes. Or at least there is an important class of physical events that in practice are explained by reference to non-physical causes, and not otherwise. The class to which I refer is the class of ostensibly voluntary human actions which, if explicable at all, are explained by reference to beliefs, desires and other mental attitudes. There is, of course, a gap between what can be explained in practice and what might be explained

in principle, but I think someone arguing "in principle" beyond the evidence of practice carries a burden of proof that the extension is justified. Anyone claiming that ostensibly voluntary actions are explicable in purely physical terms needs to argue not only that beliefs, desires and other mental attitudes are physical, but that their causal efficacy is governed by laws.

Certainly some have tried to argue this way. The first conjunct follows easily from the assumption that nothing is non-physical, so that mental states, if they exist, must be physical. The negative existential assumption that nothing is non-physical cannot aspire to be more than an assumption, unless one defines "physical" so as to make it a tautology. If the assumption be granted, it can then be asserted that although we don't yet know the applicable laws, those laws are "out there", waiting to be discovered. A bolder step is to deny the existence of mental states, and to proclaim that not only the laws, but the causally efficacious physical states governed by the undiscovered laws also await discovery by a future enlightened eliminativist science.

4.15 Conclusion

In this chapter, I have argued that there is no compelling *a priori* argument for causal closure of the physical domain, nor any empirical evidence which requires acceptance of that principle. The widespread acceptance of the principle may be partly attributable to an overly optimistic inductive argument from the past successes of physics that seemed plausible in the nineteen-fifties, but becomes less attractive as decades pass, and physics comes no closer to explaining mental phenomena like apparent free will, and Chalmers' (1996) "hard problem" of consciousness. Once the conceptual possibility is acknowledged of non-physical causal influences that do not involve forces or require energy transfer, neither the conservation laws nor the Second Law of Thermodynamics demand causal closure of the physical domain.

In Chapter 2, I provided reasons for believing that human agents have a capacity to exercise metaphysical free will, and in Chapter 3, I gave what I believe are sound arguments that the exercise of metaphysical free will is incompatible with causal closure of the physical domain. Given that incompatibility, and accepting the

reality of metaphysical free will, the physical causal closure principle must be rejected.

I have raised the conceptual possibility of non-physical causal influences that do not involve forces or require energy transfer. In the next chapter I shall show that quantum theory leaves the way open for such non-physical causal influences to affect token events in the physical domain, and on some interpretations even requires the intervention of a conscious agent. In the final chapter, I shall describe some active research projects that propose models of mind-body interaction.

CHAPTER 5

FREE WILL AND PHYSICS

5.1 Introduction

In Chapter 2, I distinguished the concept of metaphysical free will from other concepts of free will that have concerned philosophers. There are good reasons to believe that human agents actually have metaphysical free will. In the first place, it provides a simple explanation for the common intuition that some of our actions are "up to us", and secondly, it is hard to explain the evolution of consciousness unless consciousness somehow affects physical events.

In Chapter 3, I offered arguments to show that metaphysical free will is incompatible with a "causally closed" physical domain: one in which all physical events have at least their probabilities determined by prior physical events and states of affairs, and I suggested that unless there are compelling grounds to believe that the physical domain is causally closed, it is more plausible to conclude that metaphysical free will is part of the total reality.

In Chapter 4, I considered the possible sources of a widespread belief among philosophers if not scientists that the physical domain is causally closed, in the sense that every effect observed in the physical domain has its causes wholly within that domain. Although classical physics arguably entails or presumes causal closure, I argued that nothing in contemporary physics or in the empirical special sciences compels us to believe that the universe is causally closed in the sense described.

Physics alone provides no account of free will, but contemporary physics, unlike classical physics, at least does not preclude it.

If I am right in asserting that humans possess a kind of metaphysical free will that is incompatible with causal closure of the physical domain, and if there is no compelling reason to maintain that the physical domain *is* causally closed, it follows that a theory of free will must allow for entities and influences that lie outside the ontology recognised by physics. According to the folk notion of free will, it is mental events that cause voluntary actions. A theory of free will that renounces causal closure will therefore resemble an interactive dualist theory, where the non-physical influences are identified with or somehow related to the conscious experience of causing actions.

In Chapter 6, I shall discuss and criticise some contemporary accounts by scientists and philosophers who are seeking to explain consciousness and free will by renouncing causal closure, and by recognising an ontology that extends beyond the bounds of current physics. While I shall not adopt or endorse any one of these theories, I take encouragement from the existence of several research projects admitting an extended ontology. I claim that such avenues of research hold more promise than the more conventional attempts that have addressed the free will problem by redefining the concept of freedom to make it compatible with an unnecessarily restrictive view of physical reality.

In concluding, I shall offer some speculation about how or whether free will may eventually be accommodated within science. Although some form of dualism may be necessary to reconcile metaphysical free will with *current* physical theories, physics itself remains incomplete, and a completed physics may contain entities, aspects or dimensions broad enough to accommodate metaphysical free will and the consciousness that exhibits it.

5.2 Testability and Causal Closure

I claim that the exercise by humans of metaphysical free will shows that the physical domain is not causally closed, at least in any contemporary sense of the word "physical". In other words, I claim that the existence of metaphysical free will falsifies any theory predicated on causal closure of the physical. In the next chapter, I

shall examine some theories which deny that the physical domain is causally closed. Those theories will make a number of empirical claims, but not all empirical claims are falsifiable, or even testable. It is reasonable to ask whether the theories discussed in Chapter 6 meet the standard of falsifiability by which the theories they supplant are found wanting, or any more liberal standard of testability which would justify taking them seriously.

The criterion of falsifiability was introduced by Karl Popper, who rejected the orthodox characterisation of science as an inductive enterprise. Popper reminded the world of Hume's critique of induction, and he argued that scientific theories are in practice developed by a process of proposing and testing hypotheses. No general statement can be deduced from any finite number of singular statements, but a single contradictory instance can prove a general statement false. Genuine scientific theories comprise sets of general statements (or laws) whose singular consequences can be tested empirically. Though a theory can never be proven true, a theory that makes testable predictions, none of which has yet proven false, is considered a good scientific theory. A theory that makes predictions which are incapable of empirical testing, or "unfalsifiable" is considered to be unscientific[47].

Popper introduced the falsifiability criterion in his *Logik der Forschung*, published in 1934. Though it was considered radical at the time, the falsifiability criterion had become orthodox among physical scientists by the time the book was published in English (Popper 1959). I think it is fair to say that the falsifiability criterion is still endorsed within scientific circles, if not as the strict demarcation between science and "non-science", at least as a benchmark. Scientific practice routinely follows the hypothetic-deductive method, whereby the consequences of theories are predicted and tested empirically. Where results are inconsistent with predictions, revision is required, either to the theory itself or to some of the background assumptions. But contrary to Popper's strict views, not all unfalsified but falsifiable theories are regarded as equal, and scientists are naturally guided by falsified theories in framing new ones.

---

[47] Theories in general comprise a large number of general propositions, laws or statements from which, in the light of singular statements or observations, predictions can be made. A theory may be unscientific in the sense that some of its general statements allow unfalsifiable predictions, yet may be falsified as a whole because other statements make false predictions.

According to Godfrey-Smith (2003), Popper finds less support among philosophers of science than among scientists, and his popularity among the latter is due to their unawareness of his more extreme claims. That may well be true, but it possible that many scientists find his broad approach congenial, without accepting every detail of his writings. Single hypotheses can never be tested in isolation. Any recalcitrant observation might be attributable to the falsity of some background assumption or auxiliary hypothesis. And few working scientists would deny that a theory that has survived many attempts to falsify it is likely to be "closer to the truth" than one which has never been tested at all. But as one trained in science in the 1960's, I admire Popper's recognition of the centrality of the hypothetico-deductive method, his condemnation of *ad hoc* modification of theories to accommodate recalcitrant data, and his encouragement of bold hypotheses.

A more liberal demarcation between science and non-science is based on a principle of testability, rather than falsifiability. Scientific experiments are usually designed to test a chosen hypothesis by basing a prediction on that hypothesis in conjunction with other hypotheses and background assumptions that are taken by the experimenters to be true, or at least far more likely to be true than the hypothesis being tested. But in principle, a result contrary to prediction might be attributable to the falsity of any of the other hypotheses and background assumptions: a fact that any scientist would concede in principle, though not expressly consider in their everyday practice. Quine (1995: 256) explains the more liberal concept of testability as follows:

> A sentence is testable, in my liberal or holistic sense, if adding it to previously accepted sentences clinches an observation categorical that was not implied by those previous sentences alone.

An "observation categorical" in Quine's terminology is a statement asserting some regularity about the world, his example being "When it snows, it's cold".

Testability creates a problem for any theory that allows for non-physical influences. Theories are tested by making predictions that may turn out not to be true. But predictions have to be based on a knowledge of initial conditions, and on an assumption that every system with identical initial conditions will behave in a law-like way, even if those laws are intrinsically probabilistic. And that is just another

way of stating the assumption that the physical domain is causally closed. A theory that denies causal closure of the physical domain can make predictions that are not testable within that domain.

Does lack of testability mean that all interactionist theories must be rejected? I think not. I grant that the "best" theories are those that are testable but unfalsified. But if all the testable theories turn out to be false, surely an untestable theory is to be preferred over a false one. Or at best, we may have to accept some less objective[48] standard for testing the predictions of our theories. The best theories of mind might forever retain a metaphysical element.

Popper originally proposed falsifiability as a criterion for demarcation between science and other intellectual pursuits, including "speculative metaphysics". In doing so, he distinguished his position from logical positivism in at least two ways. Not only did he reject the verification principle on the ground that no empirical theory is conclusively verifiable, but he emphasised that his criterion was not one of meaning, but of demarcation (Popper 1959: 16 fn.). For Popper, an unfalsifiable theory is not thereby devoid of meaning and unworthy of consideration. It merely fails to be a theory of empirical science. As such, Popper does not condemn it "to the flames" as Hume might, but recognises it for what it is. The very principle of causality provides an example:

> [The 'principle of causality',] that the world is governed by strict laws … is *not falsifiable*. …I shall therefore neither adopt nor reject the 'principle of causality'; I shall be content simply to exclude it, as 'metaphysical' from the sphere of science. (Popper 1959: 39)

Popper expressly declines to address the question of free will in his *Logic of Scientific Discovery* (1959: 165). Nor does he discuss free will in the *Postscript* to that work, though he claims in *Quantum Theory and the Schism in Physics* that indeterminism is necessary to explain "the phenomenon of voluntary movement in animals" (1982c: 209). But in his later collaboration with Eccles, he defends an interactionist model of the mind and body that he must have recognised as unfalsifiable. His willingness to

---

[48] As to whether testing must be intersubjectively verifiable, see Section 6.5.4, below.

entertain unfalsifiable theories of mind is explicit in one of the dialogues with Eccles, concerning whether animals have "experiences just like ours":

> [T]here is no direct evidence, and I would therefore describe the problem of consciousness of animals as a kind of metaphysical problem, in the sense that any hypothesis, any conjecture about it, is not falsifiable – at any rate not at present. And since it is not falsifiable or testable, it is metaphysical.
>
> But metaphysical hypotheses are important for science in at least two ways. First of all, in order to have a general picture of the world we need metaphysical hypotheses. Secondly, in the actual preparation of our research we are guided by what I have called "metaphysical research programmes" (Popper and Eccles 1974: 442).

Quine's more liberal standard of testability admits sentences (or theories) that would be regarded as metaphysical by Popper. But Quine allows that even non-testable sentences may qualify as "good science":

> We believe many things because they fit in smoothly by analogy, or they symmetrize and simplify the overall design. Surely much history and social science is of this sort, and some hard science. Moreover, such acceptances are not idle fancy; their proliferation generates, every here and there, a hypothesis that can indeed be tested. Surely this is the major source of testable hypotheses and the growth of science. (Quine 1995: 256)

Some of the theories I shall discuss in Chapter 6 are not in all respects testable, but by Quine's criterion, they might still be "good science". If testability provides the demarcation between science and metaphysics, those theories are metaphysical but no less worthy of consideration for all that. If I venture into areas of speculative metaphysics, I do so without apology, and take encouragement from the father of falsifiability, Sir Karl Popper.

5.3 Consciousness and Free Will

In Chapter 2, I argued that human agents sometimes exercise metaphysical free will. For an agent *A* to exercise free will in that sense requires that for some event *E*, and at some time prior to *E*, the following conjunction is true:

> *A can* bring it about or ensure that *E* occurs.
> AND
> *A can* bring it about or ensure that *E* does not occur.

In each of the conjuncts, the modal verb *can* is to be understood as expressing a *realis* modality with Dynamic aspect and modality type = 'Ability', as elucidated in Section 2.2.3. This characterisation of metaphysical free will is intended to capture the idea that whether or not *E* occurs on a particular occasion is "up to *A*", and presupposes that *A* is a conscious being.

In contrast, we may imagine a simple device *D* comprising a Geiger counter and a small radioactive source, adapted to trigger event *E* if and only if a particle is emitted from the source within some specified time interval. It would be unnatural to say it was "up to *D*" whether *E* occurs, and the above conjunction would not be true if *D* were substituted as subject, because the *can* in each conjunct would not express modality type = 'Ability'. Thus, I claim the concept of metaphysical free will requires a conscious subject. In other words, consciousness is a necessary condition for metaphysical free will.

On the other hand, consciousness is not sufficient for metaphysical free will. It is quite possible imagine a hapless, conscious subject, aware of events occurring in his vicinity including the movements of his own body, but utterly incapable of controlling them[49]. Although I cannot believe that I am such a subject myself,

---

[49] Sir John Eccles argues that such a state is what normal subjects experience in dreaming:

> A characteristic feature of most dreams is that the subject of the dream feels a most disturbing impotence. He is immersed in the dream experience, but feels a frustrating inability to take any desired action. Of course he is acting in the dream,

compatibilists like Dennett (2003) and error theorists like Wegner (2002) seriously contend that all conscious agents are such subjects. They claim either that free will is an illusion, or that we have "free will" only in some sense other than the metaphysical free will described in Chapter 2.

No one, as far as I know, denies that we have consciousness. To say that consciousness is an illusion would be paradoxical, since consciousness seems to be required as the bearer of any illusion.

Thus, consciousness is necessary but not sufficient for free will. A theory (like classical physics) that fails to accommodate consciousness cannot accommodate free will. Quantum theory, in contrast, leaves the way open for consciousness. Indeed, on some interpretations, it essentially *requires* consciousness to bring about actual events. Such interpretations invite speculation that on some occasions, free choices by conscious beings affect the physical world. In this chapter, I consider some interpretations of quantum theory that accommodate consciousness and at least do not preclude metaphysical free will.

5.4 The "Minimisation of Mystery"

Here is a very bad argument:

> The mind-body problem seems intractable.
>
> The measurement problem of quantum mechanics seems intractable.
>
> Therefore, it seems that the mind-body problem and the measurement problem are one and the same.
>
> Therefore, quantum mechanics holds the solution to the mind-body problem.

> but with the experience that in doing so he is a puppet. His self-conscious mind can experience but not act effectively, which is exactly the position of the parallelists, such as identity theorists. A parallelist world would be a dream world! (Popper and Eccles 1977: 372)

David Chalmers (1995: 207) rightly parodies arguments of this sort as relying on a dubious "Law of Minimization of Mystery". Yet although its conclusion clearly does not follow from the premises as stated, that conclusion may nonetheless be true. An even worse argument would be the following:

> The previous argument is a very bad one.
>
> Therefore, quantum mechanics does *not* hold the solution to the mind-body problem.

While condemning the "Law of Minimization of Mystery", Chalmers acknowledges that quantum theory may contribute to an understanding of mental phenomena. But, if so, a lot of work remains to be done:

> Quantum phenomena have some remarkable functional properties, such as nondeterminism and nonlocality. It is natural to speculate that these properties may play some role in the explanation of cognitive functions, such as random choice and the integration of information, and this hypothesis cannot be ruled out *a priori*. But when it comes to the explanation of experience, quantum processes are in the same boat as any other. The question of why these processes should give rise to experience is entirely unanswered. (Chalmers, *loc. cit.*)

In his book, *The Conscious Mind*, Chalmers again refers to the "two great mysteries" of consciousness and quantum mechanics:

> Where there are two mysteries it is tempting to suppose that they have a common source. This temptation is magnified by the fact that the problems in quantum mechanics seem to be deeply tied to the notion of observership, crucially involving the relationship between a subject's experience and the rest of the world. (Chalmers 1996: 333)

Quite apart from any common factors between the mysteries provided by mental phenomena and quantum mechanics at its present stage of development, I think there is one compelling reason to seek an explanation for mental phenomena – and in particular free will – in terms of quantum mechanics or some extension thereof. In Chapter 1, I nailed my colours to the mast as a global realist. I believe in a single reality, mind-independent in the sense that its existence does not depend on minds, even if minds are one irreducible fragment of that reality. One important feature of

that reality – important at least for human beings – is that we have (or at least appear to have) free will. If (the appearance of) free will is to be explained, it must be explained in terms of other features of the total reality. Although I deny that physics even purports to describe the whole of reality, it offers mankind's best account of at least the non-mental features of experience, and any plausible account of free will will have to be consistent with what our "best physics" tells us about those parts of reality that fall within its ambit.

But physics remains incomplete. The classical physics of the nineteenth century works well for physical systems of everyday experience, but fails to account for the behaviour of systems at the atomic or molecular level. In consequence, it fails to account for the behaviour of macroscopic systems causally linked to discrete events at the atomic level, such as Schrödinger's (1935/1980) "diabolical contraption", or even just the needle on a Geiger counter that swings in response to a nuclear disintegration. For apparently different reasons, it also fails to account for the behaviour of systems in which velocities or the gravitational field are high in comparison to everyday terrestrial experience. Sir Roger Penrose deals informatively with the limitations of different physical theories in his book, *The Large, the Small and the Human Mind* (1997).

In the last paragraph but one, I made reference to our "best physics". The scare quotes appear because I acknowledge there is no criterion for judging what is the best physics currently available. Even if we take explanatory success as the criterion for comparing theories, different theories explain different things. Quantum field theories are most explanatory of sub-atomic, atomic and molecular phenomena, whereas general relativity is more explanatory of cosmological events and processes. Which of these is "better" depends entirely on one's field of interest. While I therefore do not claim that quantum theories are the "best" current physics in any absolute sense, I do claim that quantum theories are the best available for reconciling with free will. I make that claim because human beings (and all other life forms to which we may attribute free will) are composed of atoms, molecules and cells. Their brains and nervous systems, which seem to be intimately involved in perception and volition, operate in an environment where gravitational fields are mild, and velocities are negligible in comparison to the speed of light. Yet brains and nervous systems engage in complex chemical and electrochemical processes. The formation and

breaking of chemical bonds cannot be explained by classical physics, but is well accounted for by quantum electrodynamics (Feynman 1985).

From the standpoint of global realism, neither general relativity nor quantum theory can be true. If there is ever to be a true theory of physics, it will need to account for all physical systems, from the largest to the smallest. Its predictions would be approximated by general relativity in large systems where Planck's constant may be neglected and by quantum theory in systems where gravitational effects are insignificant, but a true theory would also have to apply to exotic systems such as black holes or the primordial universe, where spatial dimensions are small and gravity is high, and no current theory is adequate.

Among the theories of physics presently available, quantum theory holds the most promise for incorporation into a theory of free will. Classical physics holds no such promise, because it treats the physical domain as causally closed. If, as I claim in Chapter 3, metaphysical free will is incompatible with causal closure, then its explanation will not be found in classical physics. Though quantum theories are at best an approximation to the ultimate true theory of physics within a limited domain, they do not assume causal closure of physical systems. Indeed many interpretations expressly deny it. And unlike general relativity, quantum theories have demonstrated their success in accounting for events at the level of elementary particles, atoms and molecules in an environment in which life can exist. While consciousness and the nature of mental events remain mysterious to most of us, there is ample evidence from neuroscience to correlate conscious experience and mental events with chemical and electrochemical phenomena in the brain and nervous system. If mental events and consciousness are not simply physical events in the brain and nervous system, they are at least intimately associated in some way with such events. Therefore among the physical theories available, quantum theories presently offer the best prospect of reconciling metaphysical free will with physics.

5.5 Physicists and the Physical Domain

5.5.1 The Domain of Physics

In the famous 1912 presidential address to the Aristotelian Society where he likened the notion of causation to the British monarchy, Bertrand Russell also observed:

> [P]hilosophers … are too apt to take their views on science from each other, not from science. (Russell 1959: 186)

More recently, Jay Rosenberg reminded his fellow philosophers:

> [I]f Lobachevski and Riemann, Einstein and Planck, Heisenberg and Darwin have altered our understandings of ourselves and world, our philosophers' stories must perforce accommodate the lessons of Lobachevski and Riemann, Einstein and Planck, Heisenberg and Darwin. (Rosenberg 1980: xii)

As observed in Chapter 4, physicists are by no means all physicalists. When "doing physics", and thus working with phenomena in the physical domain, physicists necessarily *assume* that the domain in which they are working is closed. They deliberately design their experiments so as to exclude, as far as possible, all causal influences beyond those they can measure or control. But such an assumption does not commit them to a *metaphysical belief* in closure, any more than a theoretician working with point masses is committed to the belief that point masses are real or even metaphysically possible, or than an economist ever expects to meet the rational average man. In the sections that follow, I shall show that several of the major contributors to twentieth century physics did not share their philosopher contemporaries' predilection for physicalism.

Anyone trying to understand or apply the empirical findings of twentieth century physics is forced to give up some deeply ingrained intuitions about the physical world. Not only the behaviours we have learned from childhood, but probably the faculties evolved by countless generations of our species to interact with

the environment presuppose a structure of the physical world that is now proven to be false. For example, contrary to pre-scientific suppositions, matter is not continuous. Nor, as classical physics would have it, is matter composed of discrete elementary particles distributed through space. Well confirmed experiments such as the double slit experiment (Feynman *et. al.* 1963, Vol III: 1-4 ff.) show inescapably that the elementary particles cannot (simultaneously) be ascribed "complementary" properties such as a position and a velocity in the way that we can ascribe such properties to macroscopic entities. And that is not simply a fact about our ignorance. The intuitive model by which our species successfully interacts with the physical world, and the Newtonian physics that describes that model with precision, fail to account for physical systems at the atomic and molecular level. That failure, it must be emphasised, arises independently of any attempts to assimilate mental events into physics, or to account for free will.

Niels Bohr is supposed to have remarked that "anyone who is not shocked by quantum theory has not understood it." Introducing a series of public lectures on quantum electrodynamics, the theory for which he received the Nobel Prize, Richard Feynman says:

> [Y]ou think I'm going to explain it so you can understand it? No, you're not going to be able to understand it. Why are you going to sit here all this time, when you won't be able to understand what I am going to say? It is my task to convince you *not* to turn away because you don't understand it. You see, my physics students don't understand it either. That is because *I* don't understand it. Nobody does. (Feynman 1985: 9)

After reflecting on what is meant by the word "understand", Feynman continues:

> I'm going to describe to you how Nature is – and if you don't like it, that's going to get in the way of your understanding it. It's a problem that physicists have learned to deal with: They've learned to realize that whether they like a theory or they don't like a theory is *not* the essential question. Rather, it is whether or not the theory gives predictions that agree with experiment. (*ibid.*: 10)

In the published *Feynman Lectures on Physics*, Feynman addresses undergraduate students, among whom are future working physicists who must "learn to deal with" quantum theory. Introducing his lecture on "Quantum Behavior", he observes:

> Because atomic behavior is so unlike ordinary experience, it is very difficult to get used to and it appears peculiar and mysterious to everyone, both to novices and to the experienced physicist. Even the experts do not understand it the way they would like to, and it is perfectly reasonable that they should not, because all of direct, human experience and human intuition applies to larger objects. We know how large objects will act, but things on a small scale just do not act that way. So we have to learn about them in a sort of abstract or imaginative fashion and not by connection with our direct experience. (Feynman *et. al*. 1963, Vol III: 1-1.).

If anyone can help philosophers to look beyond the innate and ingrained assumptions that impede our theorising about Nature at the submicroscopic level, it should be a physicist who has in Feynman's words "learned to deal with it" by applying quantum-theoretical concepts in his or her mainstream work. And if only for historical purposes, it is worthwhile considering the views offered by some of the founders of quantum theory. Erwin Schrödinger's commitment to mind-body interaction has already been mentioned in Section 4.13.5. I now consider the approach of some other leading quantum theorists.

5.5.2 Bohr

Niels Bohr was the leader of the Copenhagen School, which took a pragmatic approach to quantum mechanics. According to the Copenhagen interpretation, quantum theory cannot tell us what is "really going on" between successive observations of a system, but is a superb tool for predicting the probabilities of future observations from the content of earlier observations. To speculate about the ultimate nature of physical reality is to step outside the bounds of physics. Thus, the Copenhagen interpretation makes no ontological claims or presuppositions.

> We meet here in a new light the old truth that in our description of nature the purpose is not to disclose the real essence of the phenomena but only to track down, so far as it is possible, relations between the manifold aspects of our experience. (Bohr 1934: 18)

According to Bohr, empirical quantum mechanics deals with the observation of systems by observers, who themselves remain outside the system. The observers must be treated as classical entities, and they record and communicate their observations in a language of classical concepts.

> The essentially new feature in the analysis of quantum phenomena is, however, the introduction of a *fundamental distinction between the measuring apparatus and the objects under investigation*. This is a direct consequence of the necessity of accounting for the functions of the measuring elements in purely classical terms, excluding in principle any regard to the quantum of action. On their side, the quantal features of the phenomenon are revealed in the information about the atomic objects derived from the observations. While, within the scope of classical physics, the interaction between object and apparatus can be neglected or, if necessary, compensated for, in quantum physics this interaction thus forms an inseparable part of the phenomenon. Accordingly, the unambiguous account of proper quantum phenomena must, in principle, include a description of all relevant features of the experimental arrangement. (Bohr 1958a: 310-311)

Not only must the observer of any system remain outside that system, but if we follow Bohr, an observed physical system cannot contain any other conscious or even any living being. That is because any system whose quantum mechanical evolution is to be observed must be a *closed* system, shielded from external causal influence. By their very essence, living systems sustain their own low entropy by increasing the entropy of their environment. (Schrödinger 1967: 67 ff.)

While Bohr's physics avoids speculation on the nature of unobserved systems and expressly excludes living systems from its domain, Bohr himself was happy to speculate on matters outside physics. Thus in a 1929 lecture, he speculates on how a notion of complementarity might bear upon the free will question:

> When considering the contrast between the feeling of free will, which governs the psychic life, and the apparently uninterrupted causal chain of the accompanying physiological processes, the thought has, indeed, not eluded philosophers that we may be concerned here with an unvisualizable relation of complementarity. [W]e can hardly escape the conviction that in the facts which are revealed to us by the quantum theory and lie outside the domain of our ordinary forms of perception we have acquired a means of elucidating general philosophical problems. (Bohr 1934, 100-1)

In a 1937 paper, Bohr again reflects on the nature of free will. The mere fact that causality breaks down in systems considered quantum mechanically does not explain free will, but the necessity to adopt radically new models of reality for physical systems suggests that equally radical new models may be needed to account for such phenomena of the mind:

> I am far from sharing … the widespread opinion that the recent development in the field of atomic physics could directly help us in deciding such questions as "mechanism or vitalism" and "free will or causal necessity" in favor of the one or the other alternative. Just the fact that the paradoxes of atomic physics could be solved not by a one sided attitude towards the old problem of "determinism or indeterminism," but only by examining the possibilities of observation and definition, should rather stimulate us to a renewed examination of the position in this respect in the biological and psychological problems at issue. (Bohr 1937: 295)

5.5.3 Von Neumann

Although more widely remembered for his seminal work on electronic computers, the mathematician John von Neumann also developed the first mathematically rigorous version of quantum theory in his 1932 *Mathematische Grundlagen der Quantenmechanik*, eventually translated into English in 1955. The major part of that book (von Neumann 1955), develops detailed models of quantum systems in abstract Hilbert spaces, but its final chapter offers an insightful and mathematically less daunting discussion of the measuring process.

Bohr had emphasised that within physics, it is necessary to maintain a demarcation between the prepared and observed system whose behaviour can be measured and predicted by quantum theory, and the rest of the universe, including the measuring instruments and the physicists making the measurements, described by means of classical language and concepts. In practical experimental physics the living observer and the immediate object of her sense impressions must remain outside the system, but in thought experiments one can notionally move the line of demarcation into the brain of the observer, as von Neumann discusses in 1955:

> [S]ubjective perception leads us into the intellectual inner life of the individual, which is extra-observational by its very nature (since it must be taken for granted by any conceivable observation or experiment). … Nevertheless, it is a fundamental requirement of the scientific viewpoint … that it must be possible so to describe the extra-physical process of the subjective perception as if it were in reality in the physical world – i.e., to assign to its parts equivalent physical processes in the objective environment, in ordinary space. … In a simple example, these concepts might be applied about as follows: We wish to measure a temperature. If we want, we can pursue this process numerically until we have the temperature of the environment of the mercury container of the thermometer, and then say: this temperature is measured by the thermometer. But we can carry this calculation further, and … say: [the length of the mercury column] is seen by the observer. Going still further, we could … say: [the image of the mercury column] is registered by the retina of the observer. And were our physiological knowledge more precise than it is today, we could go still further, tracing the chemical reactions which produce the impression of this image on the retina, in the optic nerve tract and in the brain, and then in the end say: these chemical changes of his brain cells are perceived by the observer. But in any case, no matter how far we calculate…, at some time we must say: and this is perceived by the observer. That is, we must always divide the world into two parts, the one being the observed system, the other the observer. … The boundary between the two is arbitrary to a very large extent. … It can be pushed arbitrarily deeply into the interior of the body of the actual observer …but this does not change the fact that in each method of description the boundary must be put somewhere, … if a comparison with experiment is to be possible. Indeed experience only makes statements of this type: an observer has made a certain (subjective) observation; and never any like this: a physical quantity has a certain value. (von Neumann 1955: 418-9)

Von Neumann goes on to develop an account of the measurement process that is later taken up by Henry Stapp, and discussed in Section 6.5 below. At this stage, I simply observe that von Neumann's interpretation requires an ineliminable observer outside any observed physical system. Either that observer is itself wholly contained within the physical domain or it is not. If the observer is *not* wholly contained within the physical domain, we are admitting the existence of a non-physical entity causally affecting the observed system, contrary to causal closure of the physical domain. If the observer *is* itself wholly contained within the physical domain, then, at least on the assumption of global realism endorsed in Section 1.1.4, it is legitimate to consider the larger system comprising the observer and the observed system. A similar dilemma

arises. If anything is to be said about that larger system, we must postulate another observer external to it. Only by denying either global realism or causal closure of the physical domain do we escape an unending regress.

Von Neumann himself may not have assented to global realism, and the passage quoted above therefore does not commit him to a denial of causal closure of the physical domain. Nonetheless, he clearly acknowledges a kind of duality: that between the observer and the system observed. Although the boundary between these two entities can be shifted, there remains an ineliminable part of the observer that can never be regarded as part of the observed system. If we regard "physical" systems as being those capable in principle of being described by physics, it seems legitimate to regard that residual part of the observer as non-physical, and to extrapolate von Neumann's duality to a dualism between the physical system and the residual, non-physical part of its observer.

5.5.4 Pauli

Another major contributor to mainstream quantum theory was Swiss physicist, Wolfgang Pauli, who worked with Bohr in the 1920's and received a Nobel prize for his discovery of the exclusion principle that bears his name. Writing in 1952, Pauli acknowledges the limited domain of physics, but looks forward to an eventual theory that will encompass both "physis and psyche":

> Since the discovery of the quantum of action, physics has gradually been forced to relinquish its proud claim to be able to understand, in principle, the *whole* world. This very circumstance, however, as a correction of earlier one-sidedness, could contain the germ of progress toward a unified conception of the entire cosmos (*Gesamtweltbild*) of which the natural sciences are only a part. (Pauli 1994: 259)

Later in the same paper, Pauli seems to deny causal closure:

> In microphysics, … the natural laws are of such a kind that every bit of knowledge gained from a measurement must be paid for by a loss of other, complementary items of knowledge. Every observation, therefore, interferes on an indeterminable scale both with the

> instruments of observation and with the system observed and interrupts the causal connection of the phenomena preceding it with those following it. This uncontrollable interaction between observer and system observed, taking place in every process of measurement, invalidates the deterministic conception of the phenomena assumed in classical physics (Pauli 1994: 260-1)

To the extent that the philosophers' faith in causal closure is based on the success of physics, it is ironic that that faith is not shared by many of the physicists themselves.

5.5.5 Heisenberg

Werner Heisenberg is most famous for his "uncertainty principle", according to which the accuracy with which mutually complementary properties of a system can be known, even in principle, is limited by Planck's constant, also called the "quantum of action". Heisenberg worked mainly in Germany, in collaboration with Bohr, and he developed the matrix formulation of quantum mechanics. In his work as a physicist, Heisenberg followed the orthodox Copenhagen approach, but throughout his life he also engaged in metaphysical reflection, as recalled in an autobiographical book, entitled *Physics and Beyond*. (Heisenberg 1971).

In a paper originally published in 1956, Heisenberg attests to the power of the Copenhagen approach on its own terms:

> The conception of the objective reality of the elementary particles has thus evaporated in a curious way, not into the fog of some new, obscure, or not yet understood reality concept, but into the transparent clarity of a mathematics that represents no longer the behavior of the elementary particles but rather our knowledge of this behavior. (Heisenberg 1958: 100)

But while the strict Copenhagen approach deals only with knowledge, Heisenberg is prepared to speculate about what happens when an observation is made, and indeed about what can be meant by the word "happens". He draws an important distinction between the physical and psychical aspects of observation:

> [O]bservation … selects of all possible events the one that has taken place. Since through the observation our knowledge of the system has changed discontinuously, its mathematical representation also has undergone the discontinuous change, and we speak of a "quantum jump". When the old adage "*Natura non fit saltus*" is used as a basis for criticism of quantum theory, we can reply that certainly our knowledge can change suddenly and this fact justifies our use of the term "quantum jump".
>
> Therefore, the transition from the "possible" to the "actual" takes place during the act of observation. If we want to describe what happens in an atomic event, we have to realize that the word 'happens' can apply only to the observation, not to the state of affairs between the two observations. It applies to the physical, not the psychical act of observation, and we may say that the transition from the "possible" to the "actual" takes place as soon as the interaction of the object with the measuring device, and therefore with the rest of the world, has come into play; it is not connected with the act of registration of the result in the mind of the observer. The discontinuous change in the probability function, however, occurs with the act of registration, because it is the discontinuous change in our knowledge in the instant of recognition that has its image in the discontinuous change in the probability function. (Heisenberg 1958a: 54-5)

Here, Heisenberg seems to envisage an irreducible duality between two processes, one of physical interaction and one of psychical observation. That in turn suggests an ontological duality between the mind of the observer and the physical systems taking part in the act of measurement. While some philosophers might dismiss the ontological views of physicists as naïve, it seems to me that those views should be treated with respect.

5.6 The Relevance of Quantum Theory

5.6.1 Quantum Indeterminism

Classical physics is deterministic, and *a fortiori* is committed to causal closure of the physical domain. But whether or not quantum theory is committed to causal closure of the physical domain depends on how one interprets it. According to the Copenhagen orthodoxy, quantum theory enables agents *outside the system being*

studied to make predictions about future observations of that system, although in general, such predictions are inherently probabilistic. The agents choose what properties of the system to observe, and those choices inevitably affect future states of the system. Although the experimental system can be shielded from outside causal influences between measurements, it *cannot* be causally closed at the time an observation is made. In a natural sense, the observation *causes* some value to be attributable to the system which could not in principle have been attributed if the agents had chosen to observe some complementary property. If it is legitimate to consider the expanded system comprising both the observed system and observer, there may be a sense in which that expanded system is causally closed. The global realism I have endorsed in Section 1.1.4 licenses the consideration of such a system, but quantum theory does not purport to account for the behaviour of such an expanded system, and offers no laws to predict the behaviour of the observer.

For example, assume a physicist prepares a system that emits a particle. Quantum theory determines the probability that the particle will be found in a given spatial region at some specified time, and it also determines the probability that the particle will be found to have a momentum within a given range at that time. But according to orthodox quantum theory, position and momentum are complementary observables, and as such, cannot both have definite values at the same time. Which of the two takes on a value at that time depends on the choice of the physicist, who is necessarily outside the system. A determined physicalist might try to account for the physicist's choice by treating her as a complex physical system, but if he attempted to apply quantum theory to the physicist as a physical system, he would need to posit some observer outside that system, and that outside observer's behaviour would be unpredictable by quantum theory. Such reasoning leads to a regression. On the globally realist assumption that the same laws apply to each successive system, we shall never be able to describe a system whose state does not depend on something external to it.

A quantum theory that relies on a separation between an observed physical system and its observer may leave the way open for metaphysical free will, but such a separation would fall far short of explaining how such free will operates. Although the presence of the observer is *required* by such a theory, the theory offers no account of that observer. The nature of the observer remains mysterious, and the theory is, on

its own terms, dualistic. It is worth stressing here that the dual ontology implicit in orthodox quantum theories does not arise from an attempt to explain metaphysical free will. Even if we assume that the observer's behaviour is wholly determined by some unknown causes, those causes would necessarily lie outside the observed physical system. The mystery of metaphysical free will is closely related to the widely discussed measurement problem intrinsic to quantum theory itself, discussed in Section 5.6.2. So to seek a solution to the free will problem through quantum mechanics is not to invoke a simplistic "law of minimisation of mystery". It is a sound strategy, recognising what seems to be a common element of irreducible mentality, both in free human actions and in the actualisation of physical systems.

Solving the measurement problem will not necessarily solve the free will problem. It may turn out that a solution to the free will problem involves concepts or entities not required to account for the behaviour of purely physical systems – if such things exist – but solving the measurement problem may at least be a step towards solving the free will problem.

5.6.2 The Measurement Problem

The measurement problem is the central problem in interpreting the mathematical model of quantum mechanics. Contrary both to ingrained intuitions and to the predictions of classical physics, quantum theory denies that the physical properties attributable to a system can all simultaneously have definite values. Any unobserved system naturally evolves into a superposition of possible states, and quantum theory accurately predicts the probability of chosen parameters being found to have values within specified ranges, if a measurement is made. The superposition is supposed to evolve deterministically with time according to Schrödinger's equation, yet superpositions are never observed.

Most interpretations of quantum theory conjecture that the act of measurement somehow causes a discontinuous "jump" from the superposed state to one in which the measured parameter has a definite value: an "eigenstate" for that parameter. Measurement is said to "collapse the wave packet" or "reduce the state function". If the system is considered as a vector in Hilbert space, measurement "projects" the state

vector onto a subspace of the unmeasured space, in which one possible value of the chosen parameter is fixed. Orthodox interpretations of quantum theory give no clear account of what constitutes a "measurement", but in some way measurement presupposes a conscious observer, who chooses which parameter to measure. Had a different parameter been chosen, a different, and possibly complementary, eigenstate would have eventuated.

5.6.3 The Dependence on Mind

The Copenhagen interpretation is pragmatically useful, in that it allows physicists to make predictions about the systems they prepare and observe. But while it requires the presence of observers outside the system, the Copenhagen interpretation expressly declines to account for the behaviour of those observers. The Copenhagen interpretation is unsatisfying as a general account of the behaviour of purely physical systems, because of its reliance on the mind of an observer to "collapse the wave packet" or "reduce the state function". It can say nothing about unobserved physical systems, yet it is difficult to accept that there is nothing that can in principle be said.

The only known life in the universe exists on or near this planet, and has existed for only a fraction of the life of the universe. Since the universe has immense numbers of galaxies with immense numbers of stars, conscious life may well have evolved in many other places, but even so, it seems that conscious life is pretty thinly distributed, and vast regions of the physical universe go unobserved. If unobserved systems merely evolve according to the Schrödinger wave function, and if large parts of the physical universe are unobserved, that seems to entail that large parts of the universe exist only as massive superpositions of states. Indeed our best cosmology tells us that the physical universe evolved through states in which temperatures were too high for even atoms to form, let alone molecules of the complexity required to support "life as we know it".

Either consciousness has some basis independent of organic life, or it somehow emerged as matter evolved into complex living systems. Many would reject the first alternative as dualistic, but if the second is true, dependence on observers to reduce superpositions implies that the universe evolved for billions of

years as a massive superposition until somewhere, on some planet, some creature evolved just sufficient consciousness to make the first observation, and precipitate the first collapse.

Strict adherence to Copenhagen principles would avoid this problem by denying it is meaningful to ask what is going on within unobserved systems. But as discussed in Section 5.5, even the founders of the Copenhagen interpretation allowed themselves to speculate beyond the bounds of physics, when not "doing their day jobs".

If a conscious observer is required to collapse the wave function, one solution would be to follow Berkeley and rely on God's mind to keep things running smoothly until mortal observers could start making their own limited observations. But if that were the explanation, it is hard to understand how physicists working in their laboratories can create and sustain some systems in superposition, and then reduce them by observation. God would have to observe everything else, but benevolently avert His gaze to avoid getting in the way of the physicists' experiments.

In one sense, the free will problem is easier than the measurement problem. If the Copenhagen interpretation is on the right track, and if some kind of mental intervention is required to reduce the wave function, then quantum theory offers nothing towards explaining the behaviour of unobserved systems. But at least the free will problem arises only for systems which are the subject of conscious awareness, so that the mental element is present, and quantum theory may contribute to the solution.

In the final chapter, I shall describe and compare various attempts by philosophers and scientists to explain consciousness in the light of quantum theory. Some of these authors expressly offer an account of free will, while others either avoid the free will question, or would even deny that humans have free will in the robust sense I seek to defend. I nevertheless find their ideas worthy of consideration. The free will problem is far from being solved, and any step towards solving the problem of consciousness may be a useful step towards solving the free will problem.

CHAPTER 6

FREE WILL WITHIN NATURE

6.1 Introduction

So far, I have argued that human agents have a real capacity to make choices that affect the physical world. That capacity, which I call metaphysical free will, is important; not only for what it tells us about the causal fabric of the world, but because its possession is tacitly presupposed in those other conceptions of human freedom that guide the social world in which people hold themselves and others responsible for their deeds.

I have further argued that the capacity for metaphysical free will cannot be explained within physics. A theory of metaphysical free will must admit that some effects in the physical domain have causes outside that domain. In the previous chapter, I showed that quite apart from the need to explain metaphysical free will, some interpretations of quantum mechanics already require some kind of influences from outside any observed physical system, or even postulate the necessity of a conscious observer to bring about events in the physical world.

In this final chapter, I begin by describing some published theories that propose an active role for consciousness in the causation of physical events. Some though not all of these theories offer an account of free will, and those that do not at least take seriously the idea that consciousness cannot be explained within physics. Given my view that the explanation of free will must lie outside physics, I am

encouraged to observe that that view is shared by others, including some physicists and brain scientists with the knowledge and resources to propose serious and specific hypotheses about how a non-physical mind might play the kind of causal role required by my concept of metaphysical free will.

Finally, in the recognition that dualist theories are not readily embraced by many philosophers, I offer some comments on the history of physics itself. I shall suggest that just as physics has advanced by entertaining collections of diverse entities and concepts until it can discover some deeper principle by which those entities and concepts are explained, it is a sound metaphysical practice to be open to a plurality of entity types, at least until those types can be united in a wider view of Nature than physics can provide.

6.2 Relative State Theories

Absent any intuitively easy solution to the measurement problem, we have either to give up, or to be willing to consider theories that are inherently unintuitive. Among such theories are the "relative state theories", which deny that state reduction ever occurs.

Such a theory was first proposed by Everett (1957), and is applied to the mind-body problem by Lockwood (1989). In essence, these theories claim that when a conscious observer "measures" a system in superposition, the state function does not collapse. Somehow or other, the consciousness of the observer "latches on" to one of the possible eigenstates, and "sees" the associated definite value, the eigenstate being a projection of the overall state relative to some preferred state of the observer's brain state which, as a whole, has become entangled with the observed system. The total system including the observer continues as a superposition of all possible outcomes, although our consciousness can never perceive the superposition as such.

The most common accounts of relative state theories are the so-called "many-worlds" accounts. According to the "many-worlds" interpretations, at every "measurement" in the history of the universe, the world splits into as many mutually inaccessible possible worlds as are required to accommodate each possible value of the measurement. Whenever a continuously variable quantity is measured, this seems to

require an uncountable infinity of worlds. Such an interpretation is explored, although not embraced, by Lewis (2004).

Arguably more parsimonious relative state interpretations are the so called "many-minds" interpretations. According to the many-minds interpretations, when a measurement is made the universe does not split, but continues to evolve as a massive superposition. It is only the consciousness of the observer that splits into a multiplicity of different minds. Although many-minds interpretations were proposed as a solution to the measurement problem of quantum mechanics, their dependence on consciousness leads Lockwood (1989, 1996) to base a theory of mind on a relative state quantum theory.

6.2.1 Lockwood

Relative state theories do not purport to *explain* consciousness, but like the standard "collapse" theories, they pre-suppose conscious minds. Relative state theories imply the existence of a universe (or universes) of unimaginable complexity or plurality, but that unimaginability is itself a feature of the theories, since any single consciousness confines itself to only one projection of the totality. Neither Everett nor Lockwood addresses the question of free choice as such, but since the minds presupposed by their theories lie outside the physical systems described by the evolving wave function, nothing within physics precludes a free choice among the available but mutually complementary observables on to which a mind attaches.

Lockwood describes the universe as a "seamless whole that evolves smoothly and deterministically in accordance with the Schrödinger equation." (1991: 228) We can consider that system as composed of a multiplicity of sub-systems, some of which correspond to spatiotemporal parts of the universe, including such things as galaxies, cats in boxes, brains, molecules and elementary particles. (There is, however, no unique or privileged decomposition, and some possible decompositions may be into components unimaginable by the human mind.) In general, each subsystem under any chosen decomposition exists as a superposition of possible states. In other words, the universe contains no definitively live or dead cats.

Lockwood does not try to explain consciousness in physical terms, but supposes that human subjects have consciousness associated with some physical subsystem of their brain. When a conscious subject observes some object subsystem of the universe, neither the object subsystem nor the subject's brain subsystem collapses to a determinate state, but the consciousness "chooses" some possible state among the superposition of states that is the brain subsystem. Relative to that possible state, there corresponds one particular state of the observed object, that shares with the brain subsystem a set of eigenstates of some preferred set of observables. Just which observables are preferred is a mystery, connected with the nature of consciousness. Though none of the objective states is any more real than any other, the conscious subject observes the object *as if* it were in the relative state designated by the consciousness:

> A conscious subject, at any given time, is entitled to think of any other subsystem, from the rest of the universe on down, as having a determinate quantum state relative to this designated state. What one would normally think of as *the* state of anything, at a given time, should really be thought of as merely its state relative to the given designated state of oneself; and this goes for the state of the universe as a whole. (Lockwood 1991: 228)

An objection often raised against relative state interpretations is that they do not account for the relative probabilities predicted by the mathematics of quantum mechanics. Thus, if there were two possible outcomes having unequal probabilities, and two equally real future worlds or consciousness streams followed the measurement, the probabilities prescribed by the Schrödinger equation would seem to have no significance. Lockwood avoids that conclusion by supposing not merely two, but an uncountable infinity of superposed brain states underlying each consciousness. A consciousness "chooses" only one of those brain those states, and through it perceives one possible state of the external world. The probability that the consciousness will find itself perceiving a state of a certain kind – say one in which the box contains a live cat – is given by the proportion of possible states where the box contains a live cat.

The theory offered by Lockwood is a theory of mind, and not a theory of free will. In fact, Lockwood expressly rejects dualism between mind and body, describing

the mind as a subsystem of the brain or body: "There is … some subset of my brain's vast number of degrees of freedom that is constitutively, and not just causally, involved in conscious mentality" (1996: 178). Lockwood's theory therefore does not allow for metaphysical free will in the sense of Section 2.2. An agent exercising metaphysical free will in that sense has the ability to *choose freely amongst* possible outcomes. The "choice" by a consciousness amongst the infinity of possible brain states of which Lockwood speaks is not a free choice, since that "choice" is determined by the probabilities of the Schrödinger equation. And neither is it a choice *among* the possible states. The agent simultaneously perceives all of the possible states:

> For it isn't as though I find myself to be in one of the eligible states *as opposed to* any of the others. On the contrary, for each of the parallel states, I have an experience of finding myself in it – though these experiences are not, of course, co-conscious. (Lockwood 1991: 229)

6.2.2 Squires

To give consciousness a *free* choice, it seems we would need to accord some privileged status to the mind that follows the chosen outcome. And that seems to require a strong form of dualism, according to which a single autonomous non-physical mind exists apart from the brain and chooses which of the equally real outcomes it observes.

A proposal somewhat along these lines is made by Squires (1988). Squires refers to Everett's original paper (1957), and notes its apparent failure to respect the probability predictions of the Schrödinger equation. On the Everett model, when an observation is made of a system of superposed states, the observer becomes entangled with the system, and the wave function then contains superpositions of states including the observer observing each possible outcome. Squires suggests that of these states, only one is *conscious*.

> We are thus saying that, in addition to the domain of *physics* – which is one world, described by a unique, deterministic wavefunction – there is something else which we can for convenience regard as *consciousness*.

> When I make a measurement my consciousness selects, at random, one of the possible outcomes. The term "at random" requires the existence of some "measure", and the only thing available here is the wavefunction, so the normal quantum probability rule is essentially a trivial consequence of the assumption. (Squires 1988: 17)

In order to employ the probabilities of the Schrödinger equation, Squires assumes determinism. He embraces a strong form of dualism, while rejecting the reality of free will that dualism offers. In order to explain why different observers agree in which outcomes they observe, he is also driven to conclude that a universal consciousness is shared by all observers.

If a theory of free will is to be built on a relative state interpretation of quantum mechanics, I think it would need to accord to each agent a non-physical consciousness, not wholly constrained by the probabilities of the Schrödinger equation but able on occasions to select *other than* randomly amongst a multitude of possible outcomes. Individual acts of free choice would respect the probability predictions of the Schrödinger equation, just as any single event cannot fail to respect a prediction that its probability of occurrence lies between 0 and 1. Probability predictions in general can be tested only within closed systems in which initial conditions can be repeatedly reproduced, and systems susceptible to external influence are, by assumption, untestable.

For me, the main attraction of the relative state interpretation of quantum mechanics is the inability of state reduction theories to account for events in unobserved systems. That seems to be one of the great problems of physics, but as free will occurs only in systems with which conscious minds are intimately involved, it may be more fruitful to search for a theory of free will within or in conjunction with a physics where minds play a role in state reduction.

6.3 The Eccles Project

6.3.1 Introduction

Australian scientist Sir John Eccles was not a physicist, but a medical scientist who earned a Nobel Prize for his seminal work on brain physiology, and in particular on the electrochemical mechanisms of synaptic transmission. In later life he recalled (Popper and Eccles 1977: 357) that it was an interest in the mind-brain problem, triggered at the age of 18, that led him to study the neural sciences and to maintain through his career "some continuing involvement in philosophy". The last two decades of his life were devoted to seeking a scientifically sound theory of mind-brain interaction, a project in which he collaborated first with philosopher Karl Popper, and later with theoretical physicist Friedrich Beck.

Eccles is an unabashed interactionist dualist. He takes the only alternative to that position to be "psycho-physical parallelism" or epiphenomenalism, a position he rejects as untenable, principally on evolutionary grounds:

> [T]here is on the parallelist view no biological reason whatsoever why the self-conscious mind should have evolved at all. If it can do nothing, what is the evolutionary meaning of it? After all, I think parallelists will agree that the self-conscious mind is in some incredible manner a result of evolution, so it has some survival value; yet it can only have survival value if it can do things. To have it in the role if a passive experience just for our enjoyment or for our suffering is an absurd notion biologically. We have to think that it was developed because of selection pressure, and so has survival value built into it. This does demand that the self-conscious mind is able to bring about changes in the brain and hence the world. (Popper and Eccles 1977: 516)

6.3.2 Testability

The model developed by Eccles and Beck combines a detailed and empirically supported physiological hypothesis about neural processes that accompany conscious events with a more speculative hypothesis about the structure and location of the non-physical mental processes themselves. Eccles readily acknowledges Popper's influence, and presents his model in ways that invite testing so far as possible. In principle, a future science of brain physiology might be able to map the structures of the brain supposed to be involved in non-deterministic choice, and to correlate activation of those structures with the agent's reports of her mental processes. But is hard to see how those mental processes, as distinct from the physical processes correlated with them, could ever be observed directly. The need to rely on the agent's reports would leave a degree of subjectivity, falling short of the standard of objective testability many would demand for hard empirical science. But even Popper (1982c: 211) admits that some metaphysical systems are worth discussing, and I would say the Eccles model is clearly one such system.

6.3.3 The Eccles-Popper Collaboration

Eccles' collaboration with Popper apparently began in New Zealand in the 1940's, when Eccles was at Otago, and Popper at Canterbury. Decades later, they co-authored a book, *The Self and its Brain*. (Popper and Eccles 1977), defending their shared belief in a form of interactive dualism, and proposing a tentative model consistent with what was then known about the neural correlates of consciousness. The book is in three parts. In the first part, Popper defends the autonomy of a non-physical self, and locates that self in the second of his three worlds, according to the ontology first described in his *Objective Knowledge* (Popper 1972). In the second part of the book, Eccles suggests a model whereby the immaterial self interacts with the brain *via* surface structures on the neo-cortex, which he calls the "liaison brain". The third part of the book is a series of recorded dialogues between the two authors

where they share and explore ideas from their perspectives as scientist and philosopher.

Speaking with authority as a neuroscientist, Eccles denies that neuroscience can account for human consciousness:

> [T[he goal of the neurosciences is to formulate a theory that can in principle provide a complete explanation of all behaviour of animals and man, including man's verbal behaviour. … With some important reservations I … share this goal in my own experimental work and believe that it is acceptable for all automatic and subconscious movements, even of the most complex kind. However, I believe that the reductionist strategy will fail in the attempt to account for the higher levels of conscious performance of the human brain. (1977: 358)

To account for this higher consciousness, Eccles conjectures that an immaterial self (a Popperian World 2 entity) scans and interacts with specific parts of the brain, the "liaison brain", tentatively identified as modular clusters of neurons (*ibid.* 366 ff.) extending down through several layers of the cortex. At any moment of consciousness, some of these clusters (Popperian World 1 entities) are "open" to influence by the self. The self performs an integrative function, "reading" patterns among these clusters, and giving rise to a global awareness.

Mind-body interaction must be a global affair, according to Eccles and Popper. They agree that the mind reads complex patterns of activity, and not the function of individual neurons: As Popper comments in one of his own chapters, "There is almost no information in the firing of one cell" (*ibid.*: 177).

In one of his dialogues with Popper, Eccles suggests that the interaction between Worlds 1 and 2 is reciprocal:

> Let us then think of the hypothesis that the self-conscious mind is not just engaged passively in reading out the operation of neural events, but that it is an actively searching operation. There is displayed or portrayed before it from instant to instant the whole of the complex neural processes, and according to attention and choice and interest or drive, it can select from this ensemble of performances in the liaison brain, searching now this and now that and blending together the results of performances in the liaison brain. In that way the self-conscious mind achieves a unity of experience.
>
> … Not only have we got the self-conscious mind actively reading out from the great display of neuronal performance in the

> liaison areas, but we also have to recognize that this activity has a feedback and that it isn't just receiving. It is also giving or acting. …
> …we might say that one small element in this feedback from the self-conscious mind to the brain is effective in bringing about mechanical events in the external world by muscles moving joints and/or causing speech and so on … (*ibid.* 472-3)

6.3.4 The Eccles - Beck Model

While Popper and Eccles were satisfied in 1977 that mind-brain interaction would not violate the conservation laws of classical physics, they recognised that the determinism of classical physics could not accommodate such interaction. Eccles went on to look for a model embracing the non-determinism of quantum theory. In the early 1990's he joined forces with Friedrich Beck, a theoretical quantum physicist, and together they developed a refinement of Eccles' earlier model, not only respecting the conservation laws, but involving physical processes in which, according to Beck's calculations, quantum uncertainties outweigh the statistical uncertainties of thermodynamics.

The model is described and defended in Beck and Eccles (1992) and Chapter 9 of Eccles (1994). A revised version was published after Eccles' death, in Beck and Eccles (2003).

In the Eccles and Beck model, the modules of interaction previously called the "liaison brain" are identified with particular features of the cerebral cortex, whose structure and functional role in perception and bodily action had been described by Szentágothai (1975). These features are constituted by bunches of a hundred or so neurons, extending through as many as five of the six layers of the neocortex. The neurons active in such structures are pyramidal cells, each having a body of roughly pyramidal shape located within one neocortical layer. The pyramidal cells are further characterised by a long apical dendrite that passes through many layers of the neocortex, and they also have numerous shorter dendrites and a single axon. Because these bunches identified by Szentágothai consist primarily of the apical dendrites of numerous functionally related neurons, Eccles (1990) adopts the name "dendrons" for them.

In common with all kinds of neurons, the pyramidal cells respond to stimuli received from other cells, and pass those stimuli on to further cells, and principally other neurons. The physiological mechanism by which neurons in general receive and transmit information is thought these days to be well understood. Eccles himself had played a major role in its elucidation before directing his research to mind-body interaction. Generally speaking, neurons have a multiplicity of relatively short, multiply-branched dendrites which can be stimulated by the receipt of a few molecules of certain amino acids, known as neurotransmitters. The neurotransmitters are passed to the dendrites by structures on the axons of neighbouring neurons across a small gap, typically about 20 nanometre (2×10^{-8} metre) wide, known as the synaptic cleft. Receipt of a dose of neurotransmitter at a synapse results in an electrical pulse being transmitted by migration of inorganic cations along the dendrite to the cell body. Pulses from a multiplicity of synaptic sites converge at the cell body, where they may or may not trigger a stronger pulse transmitted along the axon to synapses formed with other neurons.

Though highly complex, these processes can in principle be described and modelled according to classical physics. As such, they appear to be deterministic. In other words, given complete knowledge of the sequence of neurotransmitter molecules across each of a neuron's multiplicity of synapses, a Laplacian demon should be able to decide whether or not that neuron will fire within a given time interval. To discover a possible locus for significant quantum effects, Eccles looks at the finer mechanisms of neurotransmitter release.

Electron microscopy of a synapse reveals that the portion of an axon adjacent a synaptic gap takes the form of a small bulb, known as a bouton. The long apical dendrite of a typical pyramidal cell has thousands of short branches or "spines" that make synaptic contact with the boutons of other pyramidal cells. Eccles estimates there to be well over 100,000 of these synaptic connections within each dendron. Each bouton contains a solution of electrolytes, notably sodium, potassium and calcium ions. The molecules of neurotransmitter are not dissolved in the solution, but are contained within extremely small sacs (about 40 nm. in diameter) called vesicles able to float within the electrolyte. The face of the bouton adjoining the synaptic gap is a membrane formed with a dense triangular array of semi-rigid protein, the presynaptic vesicular grid (PVG), which is adapted to receive discrete vesicles, in the

manner of eggs in an egg carton. At any time, a mere forty to sixty vesicles are held in the grid, with the others floating freely in the bouton.

The process by which neurotransmitter molecules are released into the synaptic cleft is known as exocytosis. Exocytosis may be triggered when a nerve impulse travelling along the axon momentarily increases the calcium concentration in the bouton. Penetration of four calcium ions through the wall of a vesicle held in the PVG can cause that vesicle to discharge its entire contents of neurotransmitter across the synaptic cleft. Exocytosis is an all-or-nothing process. A vesicle may discharge all of its contents or none at all. At most, one of the forty to sixty vesicles on the PVG discharges in response to a single nerve pulse, a fact that Beck and Eccles speculate has to do with quantum coherence.

Empirical research cited by Eccles (1994: 152) shows that whether or not a pulse triggers an exocytosis is inherently probabilistic. Researchers who stimulated single muscular neurons with discrete pulses found probabilities of the order of 0.2 or 0.3. Where randomness is observed in nature, it may be attributable either to the statistical uncertainties of thermodynamics or to the inherent uncertainties of quantum mechanics. Which of these dominates is a function of the mass of the particle concerned and the temperature of its environment. Theoretical calculations by Beck support the view that the probabilistic uncertainty in exocytosis must be attributed, to a significant extent, to quantum mechanics.

6.3.5 Conjectures of Consciousness

The theory outlined so far is described in much more detail by Eccles and Beck in the publications already cited. The anatomical features they describe are well documented, as are the physical and electrochemical processes involved in neurotransmission. Their conjectures regarding the quantal nature of exocytosis are perhaps more controversial, but are offered in a way inviting criticism and refutation. But as theirs is a theory of interaction, Beck and Eccles must also offer conjectures about the mental entities, processes and events that by their very nature are not physical. Conjectures about non-physical entities, processes and events are not readily testable by physical means.

As noted in Section 6.3.2, it seems unavoidable that any testing of supposed correlations between non-physical mental events and physical brain processes would depend on the reports of the agent, and thus fall short of the ideal standard required by Popperians for science. But even if it is to some extent unscientific, the theory offered by Beck and Eccles offers an explanation for metaphysical free will. It may not turn out to be the "best" explanation, but if my conclusion in Chapter 3 is right, *no* theory committed to causal closure provides *any* explanation of metaphysical free will. If the only theories that are intersubjectively testable are those committed to causal closure, and if all theories committed to causal closure are false as I claim, then the best explanation obtainable must rely, in part, on a theory that is not intersubjectively testable.

Eccles supposes that all conscious experiences are composed of elemental or atomic experiences that he calls "psychons". These psychons and the rich and varied experiences constituted by them are mental entities, existing solely within Popper's World 2. Eccles further conjectures that in a living conscious brain, individual psychons interact with individual dendrons by modifying the probabilities of exocytosis at many or all of the hundred thousand or so synaptic sites on each dendron. In that way, the conscious self not only reads the subtle variations possible in each of the multiplicity of dendrons, but influences the macroscopic behaviour of the cortex , and through it the actions of the body and its physical environment.

The influence of the psychon at any one of the hundred thousand or so sites of its associated dendron may only raise the probability of exocytosis by a small amount, but the combined effect of raising the probabilities at a pattern of sites would significantly affect the pattern of impulses transmitted by neurons within the dendron.

The Eccles-Beck model has been criticised by Stapp, whose own theory is discussed below in Section 6.5. Stapp (forthcoming a) claims that by violating the basic quantum probability rules, the Eccles-Beck model would in principle allow causal anomalies, such as the possibility of sending messages backward in time. Stapp does not show how such anomalies would arise, and I would have thought that the "rules" to which he refers are empirically verified only for systems in which no mental interaction is involved. Here, we seem to be venturing into one of those areas that are untestable in principle. Eccles and Beck reject criticisms of this type on the ground that single events cannot be held accountable to probabilities:

> For the *single event* – and it has to be emphasised that quantum mechanics is a theory for the single event, and not, as is sometimes claimed, an ensemble theory – the probabilities are completely irrelevant, and the outcome is *completely unpredictable* (provided that not all but one of the probabilities are zero, which would imply the one left is equal to one). This constitutes the *non-computable* character of quantum events (Penrose 1994). It is evident that the deterministic logic underlying Cartesian dualism, which runs so heavily into conflict with the material world of classical physics, no longer applies if elementary quantum processes play a decisive role in brain dynamics. (Beck & Eccles 2003: 147)

Being neither a physicist nor a brain scientist, I hesitate to pronounce judgement on which of Stapp and Eccles and Beck is more likely to be right on this dispute. As discussed below, Stapp proposes a mechanism for interaction that does not affect probabilities, and if his criticism of Beck and Eccles on the point at issue is right, neither the possibility of mind-brain interaction in general, nor the possibility that exocytosis plays a crucial role, is ruled out.

6.4 Hodgson's Aspect Dualism

Another theory locating mind within an extension of quantum mechanics is proposed by David Hodgson (1991). Hodgson is a lawyer by profession, and his interest in the free will question is perhaps motivated by judicial considerations: How could society, and in particular how could a sentencing judge, justify penalising an individual for some act if all acts are necessitated by events beyond that person's control? Hodgson believes, as I do, that a satisfactory answer to that question cannot avoid the metaphysical question of free will. He develops his own theory of how free choice is possible in a world as described by contemporary physics.

Like Eccles (Section 6.3) and Stapp (Section 6.5), Hodgson accommodates free choice in a quantum mechanical model according to which measurement reduces the state function. His theory proposes a form of aspect dualism. Physical entities on the one hand (such as electrons, planets and brains), and mental entities on the other hand (such as conscious minds) are taken to be different aspects of a deeper reality,

represented by the quantum wave function. The mental reality does not supervene on the physical reality or *vice versa*, but each is supervenient on the underlying quantum reality. Physical events may in principle be described objectively, while mental events have a subjective character, and can only be described using the far from perfect language and concepts of folk psychology (Hodgson 1991: 381).

Hodgson's main argument for a quantum theoretical model is based on non-locality. Neuroscience tells us that representations of our environment which come through the perceptive faculties are stored in spatially separate regions of the brain, yet rational thought seems to involve the simultaneous awareness of such representations. Any instantaneous transmission of information between discrete regions of the brain would contravene special relativity, yet quantum mechanics allows for, and in fact requires, non-local correlations.

To account for free choice on Hodgson's model, one must consider both the physical and the mental aspects of the brain-mind. Consideration of the physical aspect according to quantum mechanical principles shows how a state can evolve into a superposition of states which in principle might accord probabilities to alternative developments leading to actions. When the agent acts, the actualisation of one of those possibilities appears random from the physical aspect, and from that aspect, there is no explanation why that particular alternative occurs. But the agent's mental state offers a separate, non-physical perspective of the same underlying quantum reality. From that perspective, the agent is conscious of choosing one alternative, and can sometimes even offer an explanation, couched in the language of folk psychology, for want of any better way of describing that aspect of the underlying reality. Here (*ibid.*: 392) Hodgson sees an explanation for the undeniable difference people feel between experiencing and doing. Experiencing is associated with the deterministic development of a system according to Schrödinger's equation, whereas doing involves some kind of state reduction.

6.4.1 Probabilities and State Reduction

Like the Eccles and Beck model, Hodgson's account has been criticised on the ground that Schrödinger evolution attributes precise probabilities to the alternatives and a

genuine free choice would violate those probabilities. Hodgson considers such criticisms (2005), but he does not find them damaging. For a start, the probabilities arise from a complex system which itself includes mental components in addition to those components underlying the physical reality. Thus in principle, and not only in practice, it is impossible to know what the probabilities might be. The Schrödinger equation takes no account of any non-physical features of a system, and the probabilities it prescribes are inapplicable to systems that are not wholly physical. Furthermore, each event involving a mental choice is *ipso facto* unique. Probabilities can be tested only when one can imagine identical events or states of affairs being repeated.

6.5 Stapp's Interactive Model

6.5.1 Background

Henry Stapp approaches the mind-body problem from a background in theoretical physics. Whereas Eccles sought an explanation for the efficacy of conscious events within the physiological systems and processes that were the subject of his early career, Stapp begins with a conviction that physics must acknowledge a mental aspect to reality, on which it depends for its practical application but for which it cannot account within its own laws.

Practical physics is concerned with the study and interpretation of physical systems, which it assumes to be causally closed, at least between observations. Though classical physics is now known to provide an imperfect description of even the physical part of reality, it functions perfectly well for many purposes, and assumes the physical domain to be not only closed, but deterministic. Although orthodox interpretations of quantum theory involve external causal influences when a physical system is measured, they nevertheless proceed on the assumption that the system evolves free from external causal influence between measurements. Quantum theory does not purport to account for the process of measurement, nor to explain which of a set of possible values is observed on any single occasion.

Thus, the laws and practice of physics expressly exclude mental events and entities. Although some philosophical reductionists may claim that mental events and entities are nothing over and above physical events and entities, no reduction has yet been demonstrated, and working psychologists practise their special science by reference to entities and processes whose physical realisation is at best unknown. Working physicists, in contrast, confine their investigations to systems from which mental events and entities are expressly excluded. Outside working hours – or perhaps in their retirement – some of them speculate beyond the boundaries of closed physical systems, but when they do so, they are not "doing physics". Nonetheless, I would claim that a man or woman who has trained and worked within physics is better equipped than most of us to speculate about how to reconcile mental events and processes with the physical world as revealed by empirical science. At least such a person is less likely to get the physics wrong.

Stapp had made significant contributions to mainstream quantum theory before directing his attention to questions of consciousness and free will. In a 1999 paper, he reflects on the human intuition "that our thoughts can guide our actions", and he notes that if classical physical theory were true, there would be no room left for consciousness to affect the physical world. Yet he charges that too many philosophers still approach the free will problem on the assumption that classical physics is true, or at least that it provides an adequate model for the purpose:

> The enormous success of classical physical theory during the eighteenth and nineteenth centuries has led many twentieth-century philosophers to believe that the problem with consciousness is how to explain it away. … That strategy of evasion is, to be sure, about the only course available within the strictures imposed by classical physical theory. (Stapp 1999a: 144)

As Stapp reminds the reader, classical theory is "now known to be fundamentally incorrect", and has been supplanted by quantum theory, "… which reproduces all of the empirical successes of classical physical theory, and succeeds also in every known case where the predictions of classical physical theory fail[50]". (*ibid.*: 145).

---

[50] This last claim would not be true for nineteenth century physics, some of whose failures are met by general relativity, but Stapp regards Einstein's theories as part of classical physics.

Whereas classical theory abstracts away from conscious observers, quantum theory, on most interpretations, seems to require the existence of a conscious observer. While quantum theory as such does not explain how mind and body interact, it at least provides a venue for the mental, and leaves the way open for interaction:

> [Q]uantum theory involves, basically, not just what is "out there", but also what is "in here", namely "our knowledge". Consciousness is thus introduced into contemporary orthodox physical theory, not as something *whose existence needs to be explained*, but rather as something whose detailed structure and detailed connection to brain activities needs to be further explicated. (Stapp 2004: 238)

For more than half a century, quantum theory has provided a successful mathematical model, but as discussed in Section 5.6.2, its interpretation remains problematic. Although it successfully predicts observations of isolated physical systems, the Copenhagen interpretation treats quantum mechanics as a merely pragmatic tool, making no ontological claims. But Stapp wants to go further. Even if quantum theory must on principle fail to account for the behaviour of observers, quantum theory need not be the whole of physics, and certainly should not place a limit on mankind's attempts to understand Nature:

> [T]he opinion of many physicists, including Einstein, is that the proper task of scientists is to try to construct a rational theory of nature that is not based on so small a part of the natural world as human knowledge. John Bell opined that we physicists ought to try to do better than that. The question thus arises as to what is 'really happening'. (Stapp 1999a: 148)

Stapp builds his theory of "what is really happening" on an ontology of information or elementary bits of knowledge. While orthodox quantum theory is silent about the ontology *within* the systems it describes, its 'output' is information about possible observations. As Heisenberg had remarked in 1956, the mathematics of quantum theory "… represents [with 'transparent clarity'] no longer the behavior of the elementary particles but rather our knowledge of this behavior" (Heisenberg 1958: 100). Stapp aims to go beyond quantum theory, and to give an account of a wider reality including not just the knowledge of physical systems quantum mechanics

provides, but also the mental events and entities that quantum theory on its own terms is bound to exclude. Since no material account of minds is offered by quantum theory, and since quantum theory yields only information about physical systems, Stapp contends that information is more fundamental than matter, and more universal in the sense that both mind and matter are accessible to us through information:

> The underlying commitment here is to the basic quantum principle that information is the currency of reality, not matter: the universe is an informational structure, not a substantive one. This fact is becoming ever more clear in the empirical studies of the validity of the concepts of quantum theory in the context of complex experiments with simple combinations of correlated quantum systems, and in the related development of quantum information processing. Information-based language works beautifully, but substance-based language does not work at all. (Stapp 1999a: 149)

Stapp notes that it is through the exchange of information that minds and physical systems interact in at least those processes of reality that are observed by conscious agents. But it is natural to assume that many processes are unobserved, at least by humanlike agents. Stapp's main concern is with consciousness, but to accommodate that consciousness in a broader reality, he envisages a wider concept of mentality, as will be described below.

6.5.2 Mind, Body and the Heisenberg Cut

As discussed in Section 5.5.3, it was von Neumann's insight that although quantum theory requires a notional boundary between the system observed and the observer, the location of that boundary is largely arbitrary, and it can in principle be drawn so as to treat even the observer's perceptive organs and physical brain as part of the observed system. Stapp acknowledges that insight, and considers the boundary, or "Heisenberg cut", to be drawn between the physical brain and the mental consciousness of the observer. His model is undeniably dualistic, but he defends it against the criticisms standardly raised against Cartesian dualism. Cartesian dualism "… postulates the existence of two entirely different kinds of things, but provides no

understanding of how they interact, or even can interact". In contrast, according to Stapp's model:

> …the form of the interaction between the mentally and physically described aspects of nature *is specified* in von Neumann's account of the measurement process. This account is part of a careful mathematical description of the fundamental principles of quantum theory, and of how they are to be employed in practice. The specification of the form of the interaction between the two differently described aspects is an essential part of von Neumann's formulation of quantum theory." (Stapp forthcoming b: 14-5)

Furthermore, Cartesian dualism is standardly criticised because it violates the supposed causal closure of the physical domain, but as Stapp points out, that criticism assumes classical physics. Quantum theory, as made mathematically precise and ontologically interpreted by von Neumann, shows from the outset that observed systems – necessarily standing on the "physical" side of the Heisenberg cut – *cannot be* causally closed, simply because the observer is outside the system, and the act of observation inevitably disturbs the system.

On the physical side of the Heisenberg cut is the agent's brain. That brain interacts with the rest of the physical world in ways that can in principle be explained by physics. For most purposes and again in principle, interactions between the brain and the rest of the physical world can be sufficiently approximated by classical physics and the special sciences of chemistry, biochemistry and physiology. For example, it can be assumed that the brain receives information *via* the nervous system from the sensory organs, and somehow undergoes physical changes in response, thereby forming a representation of features of the world and of the agent's relationship to that world. It can also be assumed that some changes occurring in the physical state of the brain cause events in the nervous system that in turn cause bodily movements by which events in the rest of the world are effected. All these processes by which the brain interacts with its environment can be described classically.

But if the brain itself were a classical system, all of whose states were determined by previous states of itself and its environment, no role would be left for the agent's consciousness. Contrary to the agent's experience – indeed contrary to the fact of the agent's *having any* experience – none of the agent's bodily movements would be caused by the agent. The agent would have no capacity for free will. If the

brain itself were a closed system, consciousness would not be explained by the physical states of the brain, but would seem to exist independently and superfluously, in a gratuitous dualism that even Descartes would disown.

But if we accept von Neumann's model, the agent's consciousness has an essential causal role to perform. Not just when a scientist observes a pointer in a laboratory, but whenever any conscious agent receives information from the physical world, an interaction akin to a measurement takes place. Such interactions are essential to quantum mechanical processes, and there is no shortage of empirical evidence that brains contain physical structures in which quantum effects might be significant. Eccles (Section 6.3) and Hameroff (Section 6.6 below) each suggest specific sites within the brain structure where quantum processes might play a role.

6.5.3 Interaction across the Heisenberg Cut

Broadly speaking, Stapp's account allows for reciprocal interaction across the Heisenberg cut, the boundary he notionally locates between the agent's mind and brain. As the brain responds to the external world by forming an image or representation, the mind observes that image. The observation takes place across the boundary, and must be considered quantum theoretically, and with an acknowledgment that the observing mind affects the physical state of the brain. Inasmuch as nothing in quantum physics determines either the time of an observation nor which of a plurality of observables is actually observed, Stapp surmises that the mind has a free choice that allows it to affect the brain, and through it the environment. He develops a detailed theory in his collected papers, *Mind, Matter and Quantum Mechanics* (2004). That theory extends von Neumann's interpretation as discussed in Section 5.5.3.

As described by Von Neumann (1955: 417-445), quantum theory distinguishes two processes governing an observed physical system. One of these, which he calls "process 2", governs the development of a system while not being measured or observed. Under process 2, the system develops locally and deterministically, according to Schrödinger's wave equation. Although the wave equation describes the state of the system as a function of time, it does not assign real values to each

observable quantity. Rather, it allows the calculation of the probability that any observable parameter will have a value in a specified range, if a measurement of that parameter is made at a given time. Von Neumann's other process, "process 1", concerns the act of measurement. Process 1 is neither deterministic nor local. Indeed, it *requires* the intervention of an entity from outside the system. Quantum theory cannot predict which one of a range of possibilities will be observed on any particular occasion, and a measurement conducted in one spatiotemporal region can affect the values possible at remote regions, as has since been empirically demonstrated by Aspect *et. al.* (1982).

In his extension of von Neumann's model, Stapp distinguishes three processes. Stapp's Process 2 corresponds to von Neumann's process 2, but measurement according to Stapp's model involves two distinct processes, Process 1 and Process 3. Consistently with his information ontology, Stapp explains these processes using the metaphor of questions asked and answers received:

> Process **1** is the choice on the part of the experimenter about how to act. This choice is sometimes called "The Heisenberg Choice", because Heisenberg strongly emphasized its crucial role in quantum dynamics. At the pragmatic level it is a "free choice", because it is controlled *in practice* by the conscious intentions of the experimenter/participant, and neither the Copenhagen nor von Neumann formulations provide any description of the *causal origins* of this choice, apart from the thoughts, ideas, and feelings of the agent.
>
> … Process **3** is sometimes call the "Dirac Choice". Dirac called it a "choice on the part of Nature." It can be regarded as Nature's answer to the question posed by Process **1**. This posed question might be: "Will the detecting device be found to be in the state that signifies "Yes, a detection has occurred"?" Or, "Will the Geiger counter be observed to 'fire' in accordance with the experiential conditions that define a 'Yes' response?" Each Process **3** reply must be preceded by a Process **1** question. This is because the Process **2** generates a *continuous infinity* of possible questions that cannot all be answered consistently within the mathematical framework provided by quantum theory. Process **1** specifies a set of distinct allowed possible answers such that the "Pythagoras Rule" for probabilities yields the conclusion that the probabilities for the allowed possible answers sum to unity. (Stapp forthcoming a: 21-2)

The Schrödinger equation, governing Process 2, prescribes the probabilities of different measurement outcomes (different "answers" to Process 1 "questions"), but it does not determine what observables will be measured (what "questions" will be

asked). On both Stapp's and von Neumann's models, the experimenter or participant has a free choice about which observable will be measured, and thus controls Process 1. But the participant does not control Process 3. Once it is decided to measure, say the position of a particle, Process 2 can predict only the probability that it will be observed in a certain region. Neither Schrödinger's equation nor any other known law dictates the result of a single measurement. Notwithstanding the participant's freedom to choose what will be measured, the participant does not control Process 3, and seems doomed to remain an impotent onlooker.

One way of empowering the participant would be to postulate that the participant can somehow intervene in Process 2, altering the probabilities that the Schrödinger equation would otherwise produce. Another way would be that in some manner yet to be explained, the participant has some causal influence over Process 3. Such suggestions are not inconsistent with empirical findings, since on principle, one cannot perform controlled experiments on systems that are causally open, but allowing the observer to influence Processes 2 or 3 would deny the universality of the Schrödinger equation. Stapp rejects such suggestions, and has a more elegant hypothesis, whereby the mind harnesses the so-called "quantum Zeno effect".

6.5.4 The Quantum Zeno Effect

The quantum Zeno effect is not Stapp's invention. It was described theoretically by Misra and Sudarshan (1977), and implies that for some quantum systems, there is an element of truth in the classically-suspect adage that "a watched pot never boils". In essence, if a quantum mechanical system found to be in a certain state is repeatedly and rapidly measured, the likelihood of its evolving into a different state is reduced. That is explained by the fact that systems evolve under Process 2 at a finite rate.

By way of example, consider an atom that can exist in either of two discrete energy states. If the atom is observed to be in one of those two states at time *t*, the Schrödinger equation gives the respective probabilities of its later being found in each possible state as a function of time. If the atom is first observed in the higher energy state, the probability that it will assume the lower energy state by emitting a photon increases with time. But if the energy of the atom is again observed within a

sufficiently short time interval after time *t*, it is highly probable that it will be found in the same, higher energy state as previously, since the wave function will not have changed much in a short time. If the quantum Zeno effect operates, rapid and repeated observations will tend to keep the particle in the higher energy state.

The quantum Zeno effect was proposed theoretically by Misra and Sudarshan for reasons unconnected with mind-body interaction, and was later produced experimentally, in a controlled experiment in which the transition of beryllium ions between two magnetic states is inhibited by a process akin to observation. (Itano *et. al.* 1990).

According to Stapp (1999: 159), the mind harnesses the quantum Zeno effect by forming representations of some system in the physical world – say the brain or some part of the brain – both as it is, and as the mind intends it to be at some instant in the near future. By Process 1, the mind measures whether some observable or observables within the system are within some proximity to the desired state. Process□1, consistently with the probability constraints of Process 2, delivers either an affirmative or a negative response. If the response is affirmative, the mind rapidly and repeatedly makes the same measurement. As the system will have little time to evolve far away from the last measured state, the same answer will probably be returned, and the state will be maintained. If the response is negative, the mind refrains from repeating the measurement until the system has time to evolve away from the undesired state. It can then repeat the measurement, with a significant chance of receiving an affirmative response, whereupon the system can be "locked in" to the desired state. Maintenance of the system in the desired state for an appreciable time period could then trigger some event, such as the firing of a neuron. The orchestrated firing of selected neurons could manifest itself in a macroscopic event, such as a bodily movement. Apparently the time intervals required for states to evolve sufficiently are of the order of hundreds of milliseconds. That is consistent with the observations of Libet (1985) and others that the initiation of acts of volition takes times of a similar order of magnitude.

Stapp likens this process of repeated type 1 processes to the mind's concentration on some goal. That seems plausible to me. Any intentional action such as a bodily movement requires the agent's continual awareness and monitoring of the action as it is performed. Experience suggests that more difficult actions require a

greater degree of awareness and monitoring by the actor. And goal-directed awareness and monitoring of an action seem to me to be very like concentration on that goal and its performance. Stapp finds in his theory a resonance with some ideas of William James, more than a hundred years earlier:

> The essential achievement of the will, in short, when it is most 'voluntary,' is to attend to a difficult object and hold it fast before the mind. ... Effort of attention is thus the essential phenomenon of will.
>
> …Everywhere, then, the function of effort is the same: to keep affirming and adopting the thought which, if left to itself, would slip away. (James 1890: 564 *et seq*)

Stapp follows von Neumann and Wigner in surmising that when an observation is made, the observed system generally becomes entangled with that of the measuring apparatus, which in turn becomes entangled with the mental state of the observer, unless or until the wave function (or vector) collapses into an eigenstate. Like Wigner, Stapp claims that the observer's consciousness brings about such a collapse. He acknowledges that this is an unorthodox position, but defends it vigorously:

> [Most efforts to modify the Copenhagen formulation aim] to improve it by *removing* the consciousness of the observer from quantum theory: they seek to bring quantum theory in line with the basic philosophy of the superseded classical theory, in which consciousness is imagined to be a disconnected passive witness.
>
> I see no rationale for this retrograde move. Why should we impose on our understanding of nature the condition that consciousness not be an integral part of it, or an unrealistic stricture of impotence that is belied by the deepest testimony of human experience, and is justified only by a theory now known to be fundamentally false …? (Stapp 1999a: 152, Stapp's emphasis)

6.5.5 Perception and the Body-World Schema

On Stapp's account, perception is preceded by the entanglement of the perceiver's brain with the object of perception. As some fact (such as some parameter having a value within a specified range) becomes known, the state of the perceiver's brain (and thus of the universe of which the brain is a subsystem) is thereby reduced to eliminate

the non-actualised alternatives. The acquisition of the fact is a process, physically represented as a projection operator, that reduces the brain and universe states to states compatible with the acquired information.

A conscious person has a continual awareness of his body and its immediate environment, and that awareness either is or correlates with some neural state, which Stapp dubs the person's "body-world schema". According to Stapp, a person's consciousness can also represent potential future developments of his body-world schema, including intended actions of the person's body. The representation of an intended action exists in superposition with alternatives to that action. The intention to perform some bodily action is represented by a projection operator to reduce the world-body schema to one compatible with the performance of that action, and the conscious thought actualises a pattern of brain activity which tends to bring about the action.

6.5.6 Testability and Subjectivity

Like that of Eccles, Stapp's model combines testable empirical science with speculative metaphysics. Though the quantum Zeno effect has been tested empirically in physical systems, the suggestion that it plays the role proposed by Stapp in acts of volition is untestable, at least by currently available means[51].

Though speculative, Stapp's quantum Zeno hypothesis is consistent with the impression that some acts of volition require effortful concentration on the goal to be achieved, but our lack of access to the intrinsic nature of our mental and brain processes makes it impossible to link predictions of particular observable events with objectively describable measurement processes in the brain. In particular, the conscious choice by the agent in Process 1 is *ex hypothesi* non-physical, and inaccessible to other persons. Although one could imagine the development of physiological measurement systems that might one day detect patterns of Process 3

---

[51] It is possible for an untestable and therefore metaphysical theory to become scientific as opportunities to test its predictions become available. Popper (1983a:191ff.) cites atomism as such a theory, which became testable when it made predictions about the sizes of atoms.

measurements in critical parts of the brain, any correlation of those patterns with Process 1 events would have to depend on the agent's own introspection. Though such future measurement systems might even allow an agent to test Stapp's quantum Zeno hypothesis in a purely subjective way, any observed correlations would not be intersubjectively verifiable[52].

It must be conceded that as free agents, we never seem to be aware of the brain events that Stapp surmises trigger our chosen actions, but I do not see that as damaging to Stapp's theory. It could equally be said that we are unaware of the complex processes by which our optical systems respond to stimuli of retinal nerves, and translate that information into a visual experience. Yet we see. Popper makes a similar point in his dialogues with Eccles, describing how the mind interacting with the "liaison brain" of the early Eccles model is conscious of the informational content, rather than the physical state of the brain:

> When we read a book, we very soon become quite unconscious of the letters and even the shapes of the words which we see, and the mind begins to read the meaning directly; the meaning as such. Of course, we do also read the words, but only in context and as carriers of a meaning. … In perception we read the meaning of the neuronal firing pattern of the brain and the meaning of the neuronal firing pattern is, as it were, the situation in the outside world which we try to perceive. (Popper and Eccles 1977: 478)

As human individuals grow from infancy, and as the human species evolves, it is the content of our experiences rather than the underlying processes of which we need to be conscious to survive and prosper. That seems just as applicable to experiences of acting as to experiences of perceiving.

---

[52] It is a moot point whether a theory falsifiable only by tests whose results could not be intersubjectively observed counts as "scientific". In practice, I think scientists would be more than usually sceptical about any reported experiments that could not in principle be tested by others. Intersubjectivity was part of Popper's requirement for scientific statements to be objective:

> Now I hold that scientific theories are never fully justifiable or verifiable, but that they are nonetheless testable. I shall therefore say that the *objectivity* of scientific statements lies in the fact that they can be *inter-subjectively tested*. (Popper 1959: 22)

As an agent, at any time I am conscious, I can devise simple tests such as choosing to raise my finger and doing so, or performing a purely mental act such as choosing to count backwards by sevens from ninety-one and doing so. Barring unexpected failures, the act I perform will conform to the choice I had not made before the process began. If on some occasion I choose to raise my finger and then fail to do so despite my best effort, I shall not regard the proposition that I have free will as disproved. Not unreasonably, I shall seek an explanation in terms of exceptional circumstances not present on the countless past occasions when I succeeded in performing similar acts at will.

However convincing I find such tests, they will not satisfy those philosophers who deny that I, they or anyone else has metaphysical free will. In the first place, the only evidence they can have that I chose (or seemed to choose) to act as I did is my own report. Even if my truthfulness is not challenged, I cannot disprove my opponent's alternative hypothesis that the apparent "choice" I made was in fact an illusion: that the action I took was in fact determined by circumstances of which I was not aware or occurred randomly, but that I like other members of my species have evolved to believe we are acting freely when we cannot.

Drawing on my own introspective evidence, the only argument I have is an inference to the best explanation. My belief, right now, that it is raining outside is better explained by the evidence of my senses that it is raining than by a hypothesis that I am dreaming or deluded. Similarly, I would say my belief that after starting to type this sentence I chose to stand up and turn anti-clockwise through 360 degrees – as I have just done – is better explained by my introspective evidence that I made a choice than by some hypothesis that my action was caused by circumstances of which I am not aware, or that it happened without a cause.

Unlike the Eccles and Beck model discussed in Section 6.3 and the Penrose and Hameroff model to be discussed in Section 6.6, the Stapp model has little to say about brain physiology. Because the other two models specify the brain structures in which interaction takes place, it is conceivable in principle, that a future science of brain physiology will be able to test whether activation of those structures can be correlated with mental processes of choice as reported by an agent. At its present stage of development, such testing of the Stapp model does not seem possible, even in principle.

6.5.7 Stapp's Incompatibilism

Although philosophers debate at length whether free will and determinism are compatible, Stapp is a physicist, and does not even consider the compatibilist position. In a recent paper, he takes it for granted that determinism rules out free will:

> Classical [relativistic] physics is … *deterministic*. … Consequently, according to classical theory, the complete history of the physical world *for all time* is mechanically fixed by contact interactions between tiny component parts, together with the initial conditions of the primordial universe.
>
> This result means that, according to classical physics, *you are a mechanical automaton:* your every physical action was pre-determined before you were born solely by mechanical interactions between tiny, mindless entities. Your mental aspects are *causally redundant*. (Stapp 2004: 234)

Pace Lycan (2003), it seems to me that incompatibilism is the natural view to take. Though I have no statistical evidence to support that claim, I took such an incompatibilism for granted before I started reading philosophy, and it still seems to be the view of most non-philosophers with whom I discuss the question of free will[53]. Compatibilism has become an orthodox position amongst philosophers, but I suggest that many of its adherents who do not merely accept the position on the authority of other philosophers are driven to it by a scientifically unwarranted belief that the physical universe *is* causally closed, combined with the inescapable experience that they in fact have free will in some form or other. The compatibilist's project is then to find an account of free will sufficient to satisfy those intuitions, but not inconsistent with causal closure of the physical domain. Although Stapp does not consider compatibilism at all, I have offered arguments in Chapter 3 that *metaphysical* free will, in the sense described in Chapter 2, is incompatible with causal closure of the physical world. Stapp's own work not only shows that the physical world is not

---

[53] I am pleased to note that Hodgson (2005) agrees with this assessment, as do psychologists Greene and Cohen (2004).

causally closed, but offers a theory of how mental free choices play a causal role in some physical events.

6.6 The Penrose-Hameroff Model

6.6.1 Non-Computability

An alternative proposal to account for consciousness within quantum theory is developed by English mathematician and physicist, Roger Penrose, in collaboration with American neuroscientist and anaesthesiologist, Stuart Hameroff.

The Penrose-Hameroff model is not a theory of free will, but a theory of consciousness. In his 1994 book, *Shadows of the Mind*, Penrose rejects the idea that an immaterial mind could effect collapse of the wave function (which he denotes by **'R'**), as postulated in the Eccles and Stapp models:

> In my own opinion, it is not very helpful, from the scientific point of view, to think of a dualistic 'mind' that is (logically) *external* to the body, somehow influencing the choices that seem to arise in the action of **R**. If the 'will' could somehow influence Nature's choice of alternative that occurs with **R**, then why is an experimenter not able, by the action of 'will power', to influence the result of a quantum experiment. If this were possible, then violations of the quantum probabilities would surely be rife! For myself, I cannot believe that such a picture can be close to the truth. (Penrose 1994: 350)

Nevertheless, he expressly leaves open the question of a dualistic interaction at some level or other:

> This book will not …tell us that there need necessarily be a 'self' whose actions are not attributable to external cause, but it will tell us to broaden our view as to the very nature of what a 'cause' might be. A 'cause' could be something that cannot be computed in practice or in principle. I shall argue that when a 'cause' is the effect of our conscious actions, then it must be something very subtle, certainly beyond computation, beyond chaos, and also beyond any purely random influences. Whether such a concept of 'cause' could lead us

> any closer to an understanding of the profound issue (or the 'illusion') of our free wills is a matter for the future. (*ibid.*: 36-7)
>
> [M]ight there be something that is beyond our inheritance, beyond environmental factors, and beyond chance influences – a separate 'self' that has a profound role in controlling our actions? I believe that we are very far from an answer to this question. As far as the arguments of this book go, all that I could claim with any confidence would be that whatever is indeed involved must lie in principle beyond the capabilities of those devices that we presently call 'computers'. (*ibid.*: 401)

In a jointly authored paper, Hameroff and Penrose claim as a virtue of their theory that it leaves the way open for "(apparent) non-deterministic free will", which they take to be a feature of consciousness:

> Approaches to understanding consciousness which are based on known and experimentally observed neuroscience fail to explain certain critical aspects. These include a unitary sense of binding, non-computational aspects of conscious thinking, difference and transition between pre-conscious and conscious processing, (apparent) non-deterministic free will and the essential nature of our experience. We conclude that aspects of quantum theory (e.g. quantum coherence) and of a newly proposed physical phenomenon of wave function self-collapse (objective reduction, **OR**, Penrose, 1994) offer possible solutions to each of these problematic features. (Hameroff and Penrose 1996)

The Penrose-Hameroff theory of consciousness combines Penrose's ideas about the non-computability of mental processes and his radical suggestion that gravity collapses the wave function with Hameroff's empirical research into the microstructure of brain neurons and the physiological correlates of consciousness. Penrose has long argued that consciousness cannot be explained within either classical physics or standard quantum physics, because conscious minds can operate non-algorithmically while classical and standard quantum physical systems cannot. His argument is based on Gödel's incompleteness theorem, and is set out in detail in Penrose (1994). The argument has been much criticised (for example by Grush and Churchland 1995), and defended (for example in Penrose and Hameroff 1995). I shall not enter into that debate, for whether or not the Gödelian argument is sound, I claim the existence of metaphysical free will (defended in Chapter 2 of this thesis) and its

incompatibility with causal closure (demonstrated in Chapter 3 of this thesis) provide an independent and sufficient reason to seek an explanation for consciousness outside classical physics and beyond the element of mere randomness that standard quantum theory admits.

6.6.2 Objective Reduction

One troublesome feature of standard interpretations of quantum mechanics is their apparent reliance on some kind of "measurement" to reduce the wave function. As discussed in Section 5.6.3, "measurement" seems to require some kind of conscious mind, so that no account is offered for the behaviour of physical systems with which no mind interacts. Penrose denies that consciousness is required to reduce the state of the wave function:

> Though it may well be the case that the problem of mind is ultimately related to that of quantum measurement … it is not, according to my own belief, consciousness in itself (or consciousness in the form that we are familiar with) that can resolve the internal physical issues of quantum theory. I believe that the problem of quantum measurement should be faced and solved well before we can expect to make any real headway with the issue of consciousness in terms of physical action, and that the measurement problem must be solved in entirely *physical* terms. … It is my view that solving the quantum measurement problem is a *prerequisite* for an understanding of mind and *not at all* that they are the same problem. The problem of mind is a much more difficult problem than the measurement problem! (1994: 330-1)

Penrose offers a radical solution to the quantum measurement problem, and he and Hameroff extend that solution in their theory of consciousness. In the general case of quantum systems not involving consciousness, Penrose proposes a purely physical process of "objective reduction" or **OR**, resulting from distortions in the gravitational field.

In his book *The Large, the Small and the Human Mind* (1997), Penrose describes at length the inability of present-day physics to accommodate gravity and quantum processes within a single theory. According to general relativity (itself highly successful within cosmological domains), gravity consists of a distortion or

localised curvature of space-time. Quantum theory allows for superpositions of states having different locations of matter within space-time, but treats that space-time as a rigid co-ordinate system. Major theoretical problems arise when one tries to model a superposition of states according to a relativistic concept of space-time, since the component states to be modelled by the state vector cannot be described by reference to the same co-ordinate systems. Energy differences between alternative states cannot be calculated with certainty, but are manifested in gravitational waves that are, according to Penrose, "fundamentally non-local":

> "Gravitational energy is an elusive thing. It seems to me that, if we had the right way of combining General Relativity with quantum mechanics, there would be a good chance of getting round the energy difficulties that plague theories of state-vector collapse. The thing is that, in the superposed state, you have to take into account the gravitational contribution to the energy in the superposition. But you cannot really make local sense of the energy due to gravity and so there is a basic uncertainty in the gravitational energy and that uncertainty is of the order of [the energy required to move the system from one potential state to another]. (Penrose 1997: 88-9)

Penrose rejects both relative state theories (Section 6.2) and standard interpretations of quantum mechanics in which consciousness is required to reduce the state vector, principally owing to the implausibility of vast sections of the universe existing as eternal superpositions. More satisfactory would be a theory whereby state vector collapse is a real occurrence, for which an account can be given within physics. Penrose considers the GRW approach (Ghirardi, Rimini and Weber 1986) to be too *ad hoc*., but proposes that superpositions collapse with a probability sensitive to the energy difference between possible eigenstates, attributable to differences in local curvatures of space-time.

Just as momentum and position are a pair of conjugate variables, incapable of being simultaneously attributable to a system, so are energy and time. The relationship is expressed by Heisenberg's uncertainty principle as

$$\Delta E . \Delta t = h/4\pi,$$

where E is energy, t is time, and h is Planck's constant.

Penrose conjectures that the half-life of a superposition involving uncertainty in the space-time coordinates is reciprocally related to the energy difference between the possible alternative states. The greater the distortion, the more quickly a superposition is likely to collapse to one or other eigenstate, by a process he calls "objective reduction" or "**OR**".

If Penrose is right about objective reduction, he may be a step closer to solving the measurement problem, but objective reduction as such does not offer the kind of non-computability he requires for a theory of consciousness. The instability of superpositions that distort space-time may cause a state reduction, but when such a reduction occurs, which of the possible eigenstates occurs would still be totally indeterminate. But Penrose and Hameroff claim that although the collapse of small isolated systems may be random in that way, complex structures in the brain, described in the next section, can give rise to large, spatially distributed coherent superpositions that manifest themselves as consciousness. They further propose that the collapse of these coherent superpositions can be controlled by consciousness in a process they call "orchestrated objective reduction" or "**Orch OR**":

> An important feature of **OR** (and **Orch OR**) is that non-computable aspects arise only when the quantum system becomes large enough that its state undergoes self-collapse, rather than its state collapsing because its growth forces entanglement with its environment. Because of the random nature of environment, the **OR** action resulting from growth-induced entanglement would be indistinguishable from the random **…** process of standard quantum theory.
>
> Consciousness, it is argued, requires non-computability. In standard quantum theory there is no non-computable activity, the [reduction] process being totally random. The only readily available apparent source of non-computability is **OR** (and **Orch OR**) self-collapse. An essential feature of consciousness might then be a large-scale quantum-coherent state maintained for a considerable time. **OR** (**Orch OR**) then takes place because of a sufficient mass displacement in this state, so that it indulges in a self-collapse which somehow influences or controls brain function. (Hameroff and Penrose 1996: 512*)*.

Unlike Eccles and Stapp, Penrose and Hameroff do not postulate consciousness as some kind of independent or external entity that causes reduction of the wave function. Consciousness in their model is an emergent feature of brain activity, and thus a feature of the physical domain. Nonetheless, that emergent consciousness is

presented as causally efficacious. Its collapse, in their words, "somehow influences or controls brain function". Penrose and Hameroff expressly leave open the question of whether the bearer of such consciousness really exercises "non-deterministic free will". If the appearance of such free will were merely an illusion the Penrose Hameroff model would be consistent with causal closure of the physical domain, though admitting causal influences beyond those of the Schrödinger equation.

6.6.3 Microtubules

The model described by Hameroff and Penrose accounts for the required coherent superpositions within structures whose elements are found within the microstructures of the brain. Recent research has shown that within each neuron, there exists a skeletal structure formed by tubular elements called microtubules, which apparently play a role in influencing the strength of synapses, and hence the activation of neurons by others. These microtubules have a diameter of about 25 nm, and are constructed of a protein called tubulin. The tubulin molecules are paired as dimers, the shape of each of which can readily alternate between two conformations. The massive array of these dimers, each occupying a fixed location in the tubulin wall and each capable of existing in either of two conformations is a structure potentially capable of storing digital information, comparable to 1's and 0's on a computer chip.

Hameroff's research into anaesthesia provides some empirical evidence that consciousness is associated with the activity of tubulin. To sustain coherent superpositions, the individual molecular systems would need to be somehow insulated from environmental decoherence, and Hameroff and Penrose offer hypotheses as to how that may occur. Although there is some empirical evidence for coherence in biological systems in the work of Fröhlich (1968), the model remains highly speculative.

6.7 Evaluation

6.7.1 Comparison of Theories

All of the theories presented above are to some extent metaphysical, in that they make some claims[54] that are untestable, at least intersubjectively. I have described these theories without endorsing any one of them, but I take encouragement from the existence of several avenues of investigation that take consciousness, if not free will, seriously, and try to explain mental and physical phenomena within a single world view. It would be fair to say that each of the theories embodies some features that stretch credulity in one way or another. But on further examination, it will be seen that those features are not a consequence of suppositions about non-physical causes, but arise within quantum theory itself.

The theories of Eccles, Stapp and Hodgson each rely on a supposition that mind or consciousness in some form is required to reduce the superposed state function of a physical system such as a brain, thereby altering the physical world. The difficulty that comes first to mind with such theories is that they offer no account of how purely physical events actually occur in parts of the universe not affected by mind. As mentioned in Section 5.6.3, vast regions of space and time are presumed to be devoid of life, and indeed are incompatible with the existence of physicochemical systems of the type associated with consciousness as we know it. If consciousness or mind is required to reduce the state function of physical systems in those regions of space and time, we must either assume some universal and all-pervading kind of consciousness or we must assume that the universe existed as a massive superposition for most of its history, and that unobserved systems still so exist.

While the requirement for an all-pervading consciousness is far from intuitive, the difficulty just described has nothing to do with free will. It emerges from any

---

[54] A metaphysical theory may be falsified as a whole if it contains some falsifiable elements. Thus the theory formed by conjoining "Every event has a cause" with "All swans are white" is falsified by the observation of a black swan.

attempt to give an ontological account of the Copenhagen interpretation. Stapp, for one, regards the dependence on mind as a virtue of his theory, and subscribes to an ontology of information, which he likens (Stapp 2004: 81) to Whitehead's ontology of process. The consciousness effective in human affairs is but one manifestation of a universally pervasive mental substance. For Whitehead, the most basic and fundamental components of reality are not enduring particles of matter but what he calls "actual entities" that undergo change, and typically[55] exist for only a finite duration:

> 'Actual entities' – also termed 'actual occasions' – are the final things of which the world is made up. There is no going behind actual entities to find anything more real. They differ among themselves: God is an actual entity, and so is the most trivial puff of existence in far-off empty space. But though there are gradations of importance, and diversities of function, yet in the principles which actuality exemplifies all are on the same level. The final acts are, all alike, actual entities; and these actual entities are drops of experience, complex and interdependent. (Whitehead 1929: 27-8)

Although Whitehead describes his actual entities as "drops of experience", it must be understood that there are no substantial experiencers. Actual entities are experienced or "prehended" by other actual entities, and it is relationships amongst actual entities that constitute the complex entities of the world, including those that appear as physical objects. Whitehead likens his actual entities to Leibniz's monads, but whereas Leibniz's monads are "windowless", Whitehead's actual entities essentially take cognisance of each other, and thus have an essentially mental quality.

The relative state theories of Lockwood and Squires escape the difficulty encountered by Eccles, Stapp and Hodgson by denying that reduction ever takes place. Relative state theories are also far from intuitive, since they entail a universe (or universes) of unimaginable complexity or plurality. Neither Lockwood nor Squires offers a theory of free will, but if one were prepared to entertain the dualistic concept of a non-physical mind, as Squires in fact does (Section 6.2.2), Squires' theory might be adapted by supposing that the non-physical mind has a genuine free

---

[55] Although most actual entities have a finite duration, God is the unique exception, the "primordial actual entity".

choice as to which of the multiplicity of possible states becomes entangled with the mind's consciousness. Relative state interpretations avoid the ontological commitment to process or information as more fundamental than matter, but only by accepting the reality of an unresolvable superposition of uncountable physical states. But that unintuitive commitment arises within physics, before one tries to account for free will.

6.7.2 A Hybrid Theory?

My own inclination is towards an account of physics according to which reduction of the state function really takes place, and can take place without the intervention of a conscious mind. To say that reduction *can* take place without the intervention of a conscious mind is not, of course, to deny that a conscious mind can in some circumstances govern or influence state reduction. I therefore speculate that throughout the universe, the vast preponderance of physical events involve some kind of objective reduction, perhaps as proposed by Ghirardi, Rimini and Weber (1986), or perhaps in accordance with Penrose's proposal (Section 6.6.2). But perhaps in just those rare cases where it seems that a conscious agent makes a free choice, the reduction of the wave function is influenced by some non-physical entity constituting or associated with that agent, which we identify as the agent's mind.

As noted in Section 6.6.1, Penrose himself expressly rejects the suggestion that a dualistic mind can influence the reduction of the wave function, and he intentionally refrains from addressing the free will question. My own project is more speculative, and admittedly less grounded in empirical science. Having concluded that people have a capacity for metaphysical free will, I am trying to reconcile that capacity with what is known about the physical universe, and I venture to suggest that that can be achieved only by acknowledging the potential for influence from outside the physical system. My preferred theory would therefore be a hybrid of theories which provide for objective reduction, conforming to the probabilities of standard quantum theory unless affected by a non-physical mind, associated with some of those cosmologically rare systems that exhibit consciousness. Where such minds intervene, they either effect or affect the reduction of the wave function. In these

cosmologically rare cases, reduction may be effected in the manner proposed by Stapp, harnessing the quantum Zeno effect, but as discussed in Section 6.3.5, I am not persuaded that a mind could not, in principle, over-ride the probabilities predicted by the Schrödinger equation for systems not subject to external influences[56]. However the mind influences state reduction in cases of free will, there is no need to suppose that some kind of mentality is required for events not involving conscious agents.

6.7.3 Ontological Implications

Whether or not the hybrid theory just proposed is on the right track, my arguments in earlier chapters commit me to a form of dualism. In Chapter 2, I argued that human agents have metaphysical free will in a sense that I claim in Chapter 3 is incompatible with a causally closed physical domain. Any physical theory that admits non-physical causal influences is *ipso facto* incomplete, and if I am right, any theory attempting to account for metaphysical free will in addition to the multiplicity of phenomena that current physics explains so well seems bound to be dualistic, at least in the sense of requiring two fundamentally different kinds of entity, one within physics, and one outside physics.

To a philosophical physicalist, the fact that a theory requires that kind of ontological duality may seem reason enough to dismiss it, but I disagree. I argued in Chapter 4 that neither physicalism nor causal closure can be justified by an inductive argument based on the explanatory success of physics. I shall now further argue that physics has itself entertained various ontological dualities throughout most of its

---

[56] Stapp is motivated to offer his quantum Zeno effect mechanism by his reluctance to consider that a mental interaction might interfere with the probabilities derived from the Schrödinger equation. Because the Schrödinger equation describes only the evolution of closed systems, I don't see the problem. Intervention from outside the system would not *violate* the Schrödinger equation. It would just entail that the system is not one to which the Schrödinger equation applies. For a homely analogy, assume classical physics is true, and consider a perfect billiard table, where a perfectly skilled player takes aim with a perfect cue at a perfectly spherical and elastic cue ball to sink a similarly perfect red ball. According to classical physics, the red ball will be potted. If, at the critical moment, a meddlesome child prevents the ball being potted by snatching it away, we would not say the laws of physics are thereby violated. But still less would we say that it is impossible for a child to snatch the ball because doing so would violate the laws of physics. We would simply recognise that the system of cue, balls and tables can no longer be treated as a closed system.

history, and that even in its current state of development, that part of physics open to empirical testing recognises a duality of mutually irreducible entity types. If so, then any current theory that tries to account for mental phenomena would also have to admit an ontological duality, even if mental categories are assumed to be nothing over and above physical categories. It is useful to reflect on how and why this is so, not so much as an *ad hominem* argument against the physicalist, but to indicate the direction in which an ultimate solution to the problem of free will might fruitfully be sought.

6.7.4 Dualities in Physics

The duality in current physics is not to be confused with the mind-body dualism I am prepared to entertain. Mind-body dualism as I envisage it would be an ontological duality extending beyond physics. The ontological duality *within* physics is a duality between the small and the large: between quantum theory and general relativity. Quantum theory allows successful predictions to be made of the (statistical) behaviour of small systems, in which the effect of gravity is negligible. However, its equations make no reference to gravity. General relativity is similarly successful in predicting the behaviour of large systems, where gravitational forces play a significant role, but in which Planck's "quantum of action" can be neglected. In systems where matter is highly concentrated in a small spatiotemporal region, such as black holes or the putative "big bang", neither theory is adequate, and no overarching theory has yet gained general acceptance within the community of physicists.

Through most of the twentieth century and up to the present day, physicists have sought a unified theory which could embrace the mature theories of general relativity and quantum electrodynamics, and the search continues. Although Einstein resisted Bohr's interpretation of quantum mechanics, he did not reject its empirical results. His argument in the EPR paper (Einstein *et. al.*, 1935) was that quantum mechanics as then understood was incomplete. The latter decades of his life were spent in an unsuccessful quest for a "unified field theory" encompassing electrodynamics and general relativity, as a step towards a theory which would also encompass quantum mechanics (Kaku and Thompson 1997:.130ff.).

6.7.5 Physical Methodology

Any physicalist motivated by the historical success of physics should respect the methods of physics in trying to reconcile its disparate theories. In his book *Dreams of a Final Theory,* physicist Steven Weinberg describes his approach to physics. He and two colleagues received Nobel prizes for the electroweak theory combining two of the four fundamental physical forces. Weinberg sees the goal of physics in finding theories of ever-increasing generality, able to explain an ever-expanding range of phenomena:

> We search for universal truths about nature, and, when we find them, we attempt to explain them by showing how they can be deduced from deeper truths. Think of the space of scientific principles as being filled with arrows, pointing toward each principle and away from the others by which it is explained. These arrows of explanation have already revealed a remarkable pattern: they do not form separate disconnected clumps, representing independent sciences, and they do not wander aimlessly – rather they are all connected, and if followed backward they all seem to flow from a common starting point. This starting point, to which all explanations may be traced, is what I mean by a final theory. (Weinberg (1994: 6)

Kaku and Thompson (1997) give a historical account of how physics and its predecessors in empirical science have progressively united concepts and theories previously believed to be separate.

Before Newton, it had been widely assumed that the heavens and the earth were governed by different laws:

> The laws governing the heavens were perfect and harmonious, while mortals on earth lived under physical laws that were coarse and vulgar. (*ibid*.: 18)

Quite apart from any immaterial souls, there seemed before Newton to be a duality between two kinds of material things that seemed to behave in quite different ways. Terrestrial objects like apples tended to fall towards the earth if unsupported. Stones

resisted movement if pushed, and would soon come to rest when pushing stopped. In contrast, the stars seemed to stand eternally in the celestial sphere, and the planets and moon followed incessantly the paths identified by Kepler. Newton provided new laws of mechanics and gravitation that accounted equally for the movement of the planets as well as for the behaviour of terrestrial objects. He postulated a gravitational force, acting instantaneously between any two bodies, that accounted not only for the tendency of bodies near the earth to fall downwards, but also for the maintenance of the elliptical orbits of planets and satellites. That force, directly proportional to the masses of the bodies and inversely proportional to the square of their separation, gave the right answers, although Newton could not explain its origin, and was troubled by its mysterious "action at a distance".

Newton's laws of motion and his theory of gravity eliminated one form of duality from physics, and those laws stood unchallenged for more than two centuries. It was nonetheless recognised that Newton's laws could not account for all phenomena, even in the purely physical domain. Among the phenomena left unexplained by Newton's physics were magnetism occurring naturally in some minerals, and electricity as observed in lightning and as generated by rubbing amber on cloth. By the early nineteenth century, Faraday had shown that these two phenomena, originally thought to be distinct, were somehow related, but it was not until the latter part of that century that Maxwell developed his theory of electromagnetic waves, which accounted not only for the observed electrical and magnetic phenomena but also for light, and which would survive to provide an account for then undiscovered forms of radiation such as X-rays and radio waves.

6.7.6 Symmetry in Physics

An important principle guiding twentieth century physics has been symmetry. Not only does symmetry in physical theories appeal to the aesthetic sensibilities, but in a generalised sense, symmetry has led to the development of theories which explain and predict more phenomena than the theories they displace. Symmetry in physics refers to any property of a system which is preserved under some kind of transformation. In the original geometrical sense, a figure can have rotational or reflective symmetry, but

the broader concept applies also to properties of a physical system that are invariant under permutations of its constituent elements. Such symmetries, also described as principles of invariance, led Einstein successively to his theories of special and general relativity. The symmetries that have turned out useful in the advance of physics are typically manifest only upon finding a more general way of describing a system.

Thus, Einstein's special relativity grew from the recognition that all measurable or describable properties of a mechanical system are invariant among all uniformly moving frames of reference. No uniformly moving frame is preferred over any other such frame, and in particular, the speed of light measured in any frame must be the same. Holding strictly to that principle leads to the unintuitive consequence that velocities are not additive, and in turn to the dependence of mass and linear dimensions on relative velocity, and the equivalence of mass and energy. The resultant theory of special relativity, though less intuitive than the Newtonian theory it replaced, was more symmetrical since it applied equally to all uniformly moving frames of reference.

Einstein's theory of general relativity applies not only to uniformly moving frames of reference, but to accelerating frames as well. Mutually accelerating frames are distinguishable within the terms of special relativity, but Einstein recognised that it is impossible to distinguish an accelerating frame from a uniformly moving one which is subject to a gravitational force. By expanding the theory to allow for the gravitational field, Einstein developed a theory of wider applicability: one with greater symmetry, but requiring more parameters to describe any system. The theory was also more explanatory, as gravity – a mysterious *sui generis* force under Newtonian mechanics – was now equated with the curvature of space-time itself.

6.7.7 Quantum Theories and the Standard Model

Einstein published his theory of general relativity in 1915. The theory was rapidly accepted by the community of physicists, partly because of its ability to explain the anomalous orbit of Mercury and its successful prediction of the effect of gravity on light of light waves during the 1919 solar eclipse, but no doubt also because of its

beautiful symmetry. But general relativity did not claim to be a complete theory of physics, let alone a "theory of everything". In the next decade, the theory of quantum mechanics was developed to account for a different range of phenomena.

Quantum mechanics as developed by Bohr, Schrödinger and Heisenberg was known to have a limited domain, not only omitting gravity, but neglecting even the theory of special relativity. In 1949, Richard Feynman, managed to unite quantum mechanics with special relativity, thus extending its operation into systems where velocities were not negligible in comparison to the speed of light. The resulting theory was quantum electrodynamics, or QED (Kaku and Thompson 1997: 54).

QED provides a good account of the behaviour of electrons, and thus of chemistry and the behaviour of bulk matter where gravity may be neglected. It accounts at the most fundamental level for the kinds of physical processes exhibited by living systems, including the brains and nervous systems of conscious, free human agents. But QED does not purport to explain consciousness or free will. Nor does it give any account of the systems within atomic nuclei, or of nuclear transformations. These latter physical phenomena came to be described by a further two separate theories, respectively postulating the strong and weak nuclear forces. Thus, midway through the twentieth century, physics relied on more than an ontological duality. It required four mutually irreducible theories, invoking four different fundamental forces, to account for the kinds of phenomena falling within its domain. And those four theories did not even pretend to account for consciousness or what that consciousness perceives as "free will".

Feynman received the 1965 Nobel Prize for uniting special relativity with quantum mechanics. Fourteen years later, the Nobel Prize went to Weinberg, Glashow and Salam, whose electroweak theory accounted for both the electromagnetic and the weak nuclear force and reduced the number of fundamental forces from four to three, namely the electroweak force, the strong nuclear force, and gravity. Gell Mann had received a Nobel Prize for his separate theory of Quantum Chromodynamics (QCD), accounting for nuclear particles in terms of quarks and the strong nuclear force. By the end of the 1970's, physicists had a well developed theory, the "Standard Model", which includes the strong nuclear force, as well as the electromagnetic and the weak nuclear force.

In his Nobel acceptance speech, Glashow described the new theory as "an integral work of art", and claimed that "The patchwork quilt has become a tapestry". But to this day the Standard Model still offers no account of gravity.

Within its domain, the Standard Model is highly successful. According to Brian Greene:

> During the past two decades, physicists have subjected this quantum-mechanical treatment of the three nongravitational forces – as they act among themselves [and with matter] – to an enormous amount of experimental scrutiny. The theory has met all such challenges with aplomb. (Greene 2000: 123)

Yet the failure of the Standard Model to account for the gravitational force leaves physicists unsatisfied. And although it consistently accounts for the other three forces, it lacks the simplicity and symmetry exhibited by general relativity. As Michio Kaku observes:

> At present, there has been no experimental deviation from the Standard Model. Thus it is, perhaps, the most successful theory ever proposed in the history of science. However, most physicists find the Standard Model unappealing because it is exceptionally ugly and asymmetrical. (Kaku and Thompson, 1997: 75)

The Standard Theory requires a multitude of elementary particles: thirty-six quarks, eight gluons, six leptons, a W boson, a Z boson and a Higgs boson. In addition, it requires nineteen arbitrary parameters, whose values appear to be independent, and need to be determined empirically. In spite of its success within the domain of physics where gravity can be neglected, Kaku argues such a theory must be "just an intermediate step towards a true theory of everything" . He compares the Standard Theory with theories of chemistry one hundred years earlier:

> By analogy, we can look at the Mendeleev chart, with its collection of over one hundred elements, which were the "elementary particles" of the [nineteenth] century. No one could deny that the Mendeleev chart was spectacularly successful in describing the building blocks of matter. But the fact that it was so arbitrary, with hundreds of arbitrary constants [including the respective atomic weights of the elements], was unappealing. Today, we know that this entire chart can be

explained by just three particles, the neutron, the proton and the electron (*ibid.*: 77).

6.7.8 String theory and M-Theory

At the beginning of the twenty-first century, physics has not escaped its ontological duality. Separate and seemingly inconsistent theories are required to account for gravity on the one hand, and the "Standard Model" of particles and forces on the other. And neither of these two theories has anything to say about consciousness or free will.

The residual duality *within* physics is under challenge. The most promising theories to embrace all of the physical forces, including gravity, are string theories, or a generalisation of string theories known as M-theory[57], which are described in a non-technical way by Greene (2000). According to the string theories, all of the known elementary particles and forces of physics are, at the deepest level of (physical) reality, different vibrational modes of fundamental one-dimensional strings, existing in one temporal and nine spatial dimensions. These strings are not "composed" of anything more basic, but are said to "consist" of energy. Apart from the three extended spatial dimensions of human experience, the other six dimensions have extremely high curvature, and thus escape perception[58]. Though escaping perception, the extra spatial dimensions are required by the mathematical theory, enabling all four physical forces to be treated symmetrically, and allowing the separate theories of the standard model, as well as general relativity to be derived as limiting cases. M-theory adds further symmetry by including an extra spatial dimension, admitting not just

---

[57] Since M-theory generalises one-dimensional strings to two (and higher)-dimensional membranes, the M might be thought to stand for "membrane". In fact the theory's author, Edward Witten, is coy about what the "M" signifies, suggesting in a television interview (McMaster 2003) that "M stands for magic, mystery or matrix, according to taste", or possibly even "…'murky' because our level of understanding of the theory is, in fact, so primitive."

[58] A helpful analogy is to imagine traversing the surface of a very small planet. By walking along a great circle, one would soon return to one's starting point. If the curvature of the planet were to increase, and thus its circumference to decrease to an imperceptibly small size, one could not perceive any change of location around that circumference.

one-dimensional strings, but two-dimensional membranes, and higher-dimensional "branes" within a total of ten spatial dimensions.

Notwithstanding the potency of string and M-theory to generate both general relativity and the standard model as a limiting cases, even the strongest devotees of these theories concede that they are highly speculative. There are some in-principle testable predictions consistent with those theories such as the existence of massive super-particles (Sparticles) corresponding to each elementary particle, but detection of those lies far beyond the energy range of current experimental physics. Given current experimental limitations, these theories contain much that is untestable, and at least for those sympathetic to Popper's line of demarcation, they remain metaphysical rather than physical. In a television interview, Nobel laureate Sheldon Glashow criticises string theory on just that ground:

> [S]tring theorists…have focused on questions which experiment cannot address. … I was brought up to believe, and I still believe, that physics is an experimental science. It deals with the results to experiments, or in the case of astronomy, observations.
>
> …No experiment can ever check up what's going on at the distances that are being studied. No observation can relate to these tiny distances or high energies. That is to say, there ain't no experiment that could be done, nor is there any observation that could be made, that would say, "You guys are wrong." The theory is safe, permanently safe. Is that a theory of physics or a philosophy? I ask you. (Interviewed in McMaster 2003)

Not only can we sense from Glashow's verbal style that he expects the answer "philosophy", but we have to infer that he thinks less of string theory for that reason. As a philosopher, I agree that string theory at present is more in the realm of metaphysics than of physics, but I find it no less exciting for all that.

6.7.9 Explaining Gravity Away?

So long as gravity resists incorporation into any testable theory, it seems that physicists have a choice. They can venture into metaphysical speculation like the string theorists, or they can accept at least *pro tem* the unattractive but practically serviceable duality of general relativity and the standard model of particles and forces.

Of the four forces, gravity is the odd one out. To their credit, no physical theorists have tried to escape the duality by arguing against the reality of gravity – by saying that gravity is an illusion, or otherwise attempting to explain it away. Newton admitted he could not explain the nature of gravity, but he took it into account as a real though mysterious part of nature, and he discovered laws to predict its effects. Einstein accounted for gravity in terms of space, and offered an account of gravity different from that of Newton or the contemporary folk, but each account of gravity acknowledges that the effects of gravity are real effects.

Despite the difficulty of accommodating gravity within a unified physics, no serious physicist has tried to convince us that our concept of gravity is mistaken, and that, to echo Dennett (1984), the current Standard Model of electromagnetic, strong and weak forces gives us all "the varieties of gravity worth wanting". I claim that metaphysical free will, though less widely distributed in the universe than gravity, is just as real. And therefore, until metaphysical free will can be accommodated within a testable scientific theory, it is better to accept a dualistic account of free will than to try to explain it away as something else.

While physics today is committed to an ontological duality, physicists recognise that duality as a step towards some final, unified account. Progress towards that final account is not to be made by denying the reality of those forces and phenomena which have not yet been accommodated in the theory which best accounts for the remainder.

Physics as it stands does not pretend to account for consciousness and human free agency, and there is no reason to assume that a future physics which encompasses gravity with the other three forces will account for consciousness and human free agency as serendipitous by-products. Those of us seeking to account for those phenomena would do well to follow the example of physics, accepting a plural ontology, at least until some deeper account of reality is found.

6.8 Facing up to Dualism

The position I have reached admits an ontological dualism in which non-physical minds are able to affect events in the physical domain. Though dualism is not a

popular position in contemporary metaphysics, mine is a position I am willing to defend. It is one that follows from my assumption of global realism, in conjunction with the reality of metaphysical free will as I have characterised it in Chapter 2. My arguments in Chapter 3 show that metaphysical free will is incompatible with causal closure of the physical domain, and Chapter 4 shows we have less reason to believe in physical causal closure than in the reality of our free will. Chapters 5 and 6 suggest ways in which metaphysical free will might be reconciled with contemporary physical theories, though not assimilated within them.

Though I believe metaphysical free will cannot be explained within physics, I do not equate physics with Nature, and I take metaphysical free will to be a feature of Nature. Though the best account of free will presently available may have to be dualistic, the future may hold a theory that is not. The most satisfying solution to the free will problem would one that accommodates both physical and mental phenomena within a wider view of "Nature": something like the *Gesamtweltbild* or "unified conception of the entire cosmos" to which Pauli looks forward in the passage quoted in Section 5.5.4. We are far from finding that wider conception of Nature, but if it is ever to be found, I believe the best hope of finding it will be through the methodology of the natural sciences, informed and guided by a metaphysics that can rule out some of the futile lines of enquiry.

If I have made any contribution to the project, I hope it is by helping to clear away two misconceptions: firstly that any conception of free will that does not permit a choice between alternative futures is adequate to account for the role of human agents within the world, and secondly, that the physical domain is causally closed.

BIBLIOGRAPHY

Armstrong, D.M. 1968. *A Materialist Theory of the Mind,* London: Routledge and Kegan Paul.

Aspect, Alain, Dalibard, Jean, and Roger, Gérard 1982. "Experimental Test of Bell's Inequalities Using Time-Varying Analyzers." *Physical Review Letters* **49**, 1804-07.

Austin, J. L. 1961. "Ifs and Cans." In *Austin, Philosophical Papers*, edited by J.O. Urmson and G. Warnock, 153-80. Oxford: Clarendon Press.

Baker, Lynne Rudder 1993. "Metaphysics and Mental Causation." In *Mental Causation*, edited by John Heil and Alfred Mele. Oxford: Clarendon Press.

Beck, Friedrich, and Eccles, John C. 1992. "Quantum Aspects of Brain Activity and the Role of Consciousness." *Proceedings of the National Academy of Sciences* **89**, 11357-61.

——— 2003. "Quantum Processes in the Brain." In *Neural Basis of Consciousness*, edited by Naoyuki Osaka. Philadelphia, PA: John Benjamins.

Bohr, Niels 1934. *Atomic Theory and the Description of Nature,* Cambridge: Cambridge University Press.

——— 1937. "Causality and Complementarity." *Philosophy of Science* **4**, 289-98.

——— 1958a. "Quantum Physics and Philosophy." In *Philosophy in Mid-Century*, edited by Walther Klibanski, 308-14. Florence: La Nuova Italia.

Brillouin, L. 1949. "Life, Thermodynamics and Cybernetics." *Am. Sci.* **37**, 554-68.

——— 1951. "Maxwell's Demon Cannot Operate: Information and Entropy. I." *Journal of Applied Physics* **22**, 334-7.

Broad, C. D. 1925. *The Mind and Its Place in Nature,* London: Routledge & Kegan Paul.

Campbell, C.A. 1951. "Is 'Freewill' a Pseudo-Problem?" *Mind* **LX**, 441-65.

——— 1963. "A Reply to Professor Smart's "Free-Will, Praise and Blame"." *Mind* **LXXII**, 400-05.

Cartwright, Nancy 1999. *The Dappled World: A Study of the Boundaries of Science,* Cambridge: Cambridge University Press.

Chalmers, David J. 1995. "Facing up to the Problem of Consciousness." *Journal of Consciousness Studies* **2**, 200-19.

——— 1996. *The Conscious Mind : In Search of a Fundamental Theory,* New York: Oxford University Press.

Chisholm, R. M. 1964. "J L Austin's Philosophical Papers." *Mind*, 73 1-26.

Coates, Jennifer 1983. *The Semantics of Modal Auxiliaries,* London and Canberra: Croom Helm.

Compton, Arthur H. 1935. *The Freedom of Man,* New Haven: Yale University Press.

Corrado, Michael 1991. "Criminal Law: Notes on the Structure of a Theory of Excuses." *Journal of Criminal Law and Criminology* **82**, 464-97.

Crane, Tim, and Mellor, D.H. 1990. "There Is No Question of Physicalism." *Mind* **99**, 185-206.

"Criminal Code (Queensland)." **1899.**

Crystal, David 1995. *The Cambridge Encyclopaedia of the English Language,* Cambridge: Cambridge University Press.

Davidson, Donald 1973. *Freedom to Act" in "Essays on Freedom of Action", Ted Honderich (Ed), 137-156,* London: Routledge and Kegan Paul.

Dennett, Daniel C. 1971. "Intentional Systems." *The Journal of Philosophy* **68**, 87-106.

——— 1984. *Elbow Room: The Varieties of Free Will Worth Wanting,* Cambridge, MA: Bradford Books.

——— 1984a. "I Could Not Have Done Otherwise – So What?" *The Journal of Philosophy* **85**, 553--65.

——— 1991. *Consciousness Explained,* London: Penguin Books.

——— 2003. *Freedom Evolves,* New York: Viking.

Descartes, René 1911. *Meditations on First Philosophy.* Translated by Elizabeth S. Haldane Cambridge: Cambridge University Press.

——— 1988. "The Passions of the Soul." In *Descartes: Selected Philosophical Writings*, edited by R. Stoothoff and D. Murdoch J. Cottingham. Cambridge: Cambridge University Press.

Dowe, Phil 1995. "What's Right and What's Wrong with Transference Theories." *Erkenntnis* **42**, 363-74.

Eccles, John C. 1990. "A Unitary Hypothesis of Mind--Brain Interaction in the Cerebral Cortex." *Proceedings of the Royal Society of London. Series B, Biological Sciences* **240**, 433-51.

——— 1994. *How the Self Controls Its Brain,* Berlin: Springer-Verlag.

Eddington, Arthur S. 1933. *The Nature of the Physical World,* Cambridge: Cambridge University Press.

Ehrenberg, W 1967. "Maxwell's Demon." *Scientific American* **217**, 103-10.

Einstein, A., Podolsky, B., and Rosen, N. 1935. "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" *Physical Review* **47**, 777 - 80.

Evans, Denis J., Cohen, E.G.D., and Morriss, G.P. 1993. "Probability of Second Law Violations in Shearing Steady States." *Physical Review Letters* **71**, 2401-5.

Everett, Hugh III 1957. ""Relative State" Formulation of Quantum Mechanics." *Reviews of Modern Physics* **29**, 454–62 URL= http://prola.aps.org/abstract/RMP/v29/i3/p454_1.

Feigl, Herbert 1958. "The 'Mental' and the 'Physical'." In *Minnesota Studies in the Philosophy of Science Ii*, edited by Grover Maxwell and Michael Scriven Herbert Feigl. Minneapolis: University of Minnesota Press.

Feynman, Richard P. 1985. *Q E D: The Strange Theory of Light and Matter,* Princeton: Princeton University Press.

Feynman, Richard P., Leighton, Robert B., and Sands, Matthew 1963. *Feynman Lectures on Physics*. Vol. 1.

Fowler, H.W. 1965. *Modern English Usage, Second Edition Revised by Sir Ernest Gower.,* Oxford: Oxford University Press.

Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* **66**, 66 829-39.

Fröhlich, Herbert 1968. "Long-Range Coherence and Energy Storage in Biological Systems." *International Journal of Quantum Chemistry* **2**, 641-49.

Ghirardi, G.C., Rimini, A., and Weber, T. 1986. "Unified Dynamics for Microscopic and Macroscopic Systems." *Physical Review D* **34**, 470-91.

Godfrey-Smith, Peter 2003. *Theory and Reality,* Chicago: University of Chicago Press.

Goldman, Alvin I. 1970. *A Theory of Human Action,* Englewood Cliffs, N.J.: Prentice-Hall.

Gould, S. J., and Lewontin, R. C. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London. Series B, Biological Sciences* **205**, 581-98.

Greene, Brian R 2000. *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory,* London: Vintage Books.

Greene, Joshua, and Cohen, Jonathan 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions of The Royal Society B.* **359**, 1775-85.

Grice, H.P. 1975. "Logic and Conversation." In *Syntax and Semantics, Vol. 3*, edited by P. Cole and J.L. Morgan, 41-58. New York: Academic Press.

Grush, Rick, and Churchland, Patricia S. 1995. "Gaps in Penrose's Toilings." *Journal of Consciousness Studies* **2**, 10-29.

Halliday, M.A.K, and Matthiessen, Christian 2004. *An Introduction to Functional Grammar,* London: Arnold.

Hameroff, Stuart R., and Penrose, Roger 1996. "Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: A Model for Consciousness." In *Towards a Science of Consciousness: The First Tucson Discussions and Debates*, edited by Stuart R. Hameroff, Alfred W. Kazniak and Alwyn C. Scott. Cambridge, MA: MIT Press.

Harsh, Wayne 1968. *The Subjunctive in English,* Birmingham, Alabama: University of Alabama Press.

Hawthorne, John 2001. "Freedom in Context." *Philosophical Studies*, 104(1) 63-79.

Heisenberg, Werner 1958. "The Representation of Nature in Contemporary Physics." *Daedalus* **87**, 95-108.

——— 1958a. *Physics and Philosophy: The Revolution in Modern Science,* New York: Harper & Row.

——— 1971. *Physics and Beyond*. Translated by Arnold J. Pomerans London: George Allen and Unwin.

Hempel, Carl G. 1980. "Comments on Goodman's *Ways of Worldmaking*." *Synthese* **45**, 193-4.

Hobart, R.E. 1934. "Free Will as Involving Determinism and Inconceivable without It." *Mind* **43**, 1-27.

Hodgson, David 1991. *The Mind Matters : Consciousness and Choice in a Quantum World,* Oxford: Clarendon Press.

——— 2005. "A Plain Person's Free Will." *Journal of Consciousness Studies* **12**, 3-19.

Honderich, Ted 1995. *The Oxford Companion to Philosophy,* Oxford: Oxford University Press.

Honoré, A.M.M. 1964. "Can and Can't." *Mind* **73**, 463-79.

Hornsby, Jennifer 1980a. "Arm Raising and Arm Rising." *Philosophy*, 55 73-84.

——— 1980b. *Actions,* London: Routledge and Kegan Paul.

Itano, Wayne M, Heinzen, D.J., Bolinger, J.J., and Weinland, D.J. 1990. "Quantum Zeno Effect." *Physical Review A* **41**, 2295-300.

Jackson, Frank, and Pettit, Philip 1990. "Program Explanation: A General Perspective." *Analysis* **50**, 107-17.

James, William 1879. "Are We Automata?" *Mind* **4**, 1-22.

——— 1890. *The Principles of Psychology,* New York: H. Holt and Company.

Kaku, Michio, and Thompson, Jennifer 1997. *Beyond Einstein: The Cosmic Quest for the Theory of the Universe,* Oxford, New York: Oxford University Press.

Kane, Robert 1996. *The Significance of Free Will,* New York and Oxford: Oxford University Press.

Kim, Jaegwon 1999. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation,* Cambridge Massachusetts: The MIT Press.

Kirkpatrick, Betty, ed. 1998. *Roget's Thesaurus of English Words and Phrases*. Harmondsworth: Penguin Books.

Kratzer, Angelika 1977. "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* **1**, 337-55.

Leech, Geoffrey N. 1987. *Meaning and the English Verb*. Second ed London and New York: Longman.

Leff, Harvey S., and Rex, Andrew F. 1990. *Maxwell's Demon: Entropy, Information, Computing,* Bristol: Adam Hilger.

Lewis, David K. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* **63**, 17-25.

——— 1986b. "Are We Free to Break the Laws?" In *Philosophical Papers, Vol. 2*, edited by D.K. Lewis, 291-98. New York, Oxford: Oxford University Press.

——— 1996. "Elusive Knowledge." *Australasian Journal of Philosophy* **74**, 549-67.

——— 2000. "Causation as Influence." *Journal of Philosophy* **XCVII**, 182-97.

——— 2004. "How Many Lives Has Schrödinger's Cat?" *Australasian Journal of Philosophy* **82**, 3-22.

Libet, Benjamin 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *The Behavioral and Brain Sciences* **8**, 529-39.

Locke, John 1975. *An Essay Concerning Human Understanding*. Edited by P.H. Nidditch Oxford: Oxford University Press.

Lockwood, Michael 1991. *Mind, Brain and the Quantum: The Compound 'I',* Oxford: Blackwell.

——— 1996. "'Many Minds' Interpretations of Quantum Mechanics." *British Journal for the Philosophy of Science* **47**, 159-88.

Lycan, William G. 2003. "Free Will and the Burden of Proof." In *Minds and Persons*, edited by Anthony O'Hear. Cambridge: Cambridge University Press.

Lyons, John 1977. *Semantics,* Cambridge: Cambridge University Press.

Marshall, Alan 1955. *I Can Jump Puddles.* 2nd. ed Melbourne: Lloyd O'Neil Pty. Ltd.

Matthiessen, Christian 1996. "Tense in English Seen through Systemic-Functional Theory." In *Meaning and Form: Systemic Functional Interpretations*, edited by C Butler M. Berry, R Fawcett and G Huang (eds.). Norwood N.J.: Ablex Publishing Company.

Maxwell, James Clerk 1871. *Theory of Heat,* London: Longmans, Green and Co.

McCall, Storrs 1994. *A Model of the Universe: Space-Time, Probability, and Decision,* Oxford: Clarendon Press.

McMaster, Joseph 2003. "String's the Thing." In *The Elegant Universe*. U.S.A.: A NOVA Production for WGBH/Boston and Channel 4 in association with

David Hickman Films, Sveriges Television and Norddeutscher Rundfunk, transcript available at http://www.pbs.org/wgbh/nova/transcripts/3013_elegant.html.

Melnyk, Andrew 2003. "Some Evidence for Physicalism." In *Physicalism and Mental Causation*, edited by Sven Walter and Hans-Dieter Heckmann, 155-72. Exeter: Imprint Academic.

Menzies, Peter 2003. "The Causal Efficacy of Mental States." In *Physicalism and Mental Causation*, edited by Sven Walter and Hans-Dieter Heckmann. Exeter: Imprint Academic.

Misra, B., and Sudarshan, E.C.G 1977. "The Zeno's Paradox in Quantum Theory." *Journal of Mathematical Physics* **18**, 756-63.

Mohrhoff, U 1999. "The Physics of Interactionism." *Journal of Consciousness Studies* **6**, 165-84.

Montero, Barbara 1999. "The Body Problem." *Nous* **33**, 183-200.

——— 2003. "Varieties of Causal Closure." In *Physicalism and Mental Causation*, edited by Sven Walter and Hans-Dieter Heckmann, 173-87. Exeter: Imprint Academic.

Moore, G. E. 1912. *Ethics,* London: Oxford University Press.

Morowitz, Harold J. 1987. "The Mind Body Problem and the Second Law of Thermodynamics." *Biology and Philosophy*, 2 271-75.

Nagel, Thomas 1986. *The View from Nowhere,* New York: Oxford University Press.

O'Shaughnessy, Brian 1973. "Trying (as the Mental "Pineal Gland")." *Journal of Philosophy* **70**, 365-86.

Oppenheim, Paul and Putnam, Hilary 1958. "Unity of Science as a Working Hypothesis." In *Minnesota Studies in the Philosophy of Science Vol. I I*, edited by Grover Maxwell and Michael Scriven Herbert Feigl. Minneapolis: University of Minnesota Press.

Palmer, F. R. 1990. *Modality and the English Modals.* 2nd ed London, New York: Longman.

Papineau, David 1993. *Philosophical Naturalism,* Oxford and Cambridge MA: Blackwell.

——— 2001. "The Rise of Physicalism." In *Physicalism and Its Discontents, Gillett, Carl (Ed), 3 36.* Cambridge: Cambridge Univ Pr.

——— 2002. *Thinking About Consciousness,* Oxford: Oxford University Press.

Pauli, Wolfgang 1994. *Writings on Physics and Philosophy*. Edited by C.P. Enz and K. von Meyenn Berlin: Springer.

Penrose, R., and Hameroff, S 1995. "What `Gaps'? Reply to Grush and Churchland." *Journal of Consciousness Studies* **2**, 98-111.

Penrose, Roger 1987. "Quantum Physics and Conscious Thought." In *Quantum Implications: Essays in Order of David Bohm*, edited by B.J. Hiley and F. David Peat, 105-19. New York: Methuen.

——— 1994. *Shadows of the Mind: An Approach to the Missing Science of Consciousness,* Oxford: Oxford University Press.

——— 1997. *The Large, the Small and the Human Mind,* Cambridge: Cambridge University Press.

——— 2004. *The Road to Reality: A Complete Guide to the Physical Universe*□□ITY??: Jonathan Cape.

Popper, Karl R. 1959. *The Logic of Scientific Discovery,* London: Hutchinson.

——— 1966. *The Open Society and Its Enemies*. 5th ed London: Routledge & Kegan Paul.

——— 1972. *Objective Knowledge; an Evolutionary Approach,* Oxford: Clarendon Press.

——— 1982b. *The Open Universe: An Argument for Indeterminism: From the Postscript to the Logic of Scientific Discovery,* London: Hutchinson.

——— 1982c. *Quantum Theory and the Schism in Physics: From the Postscript to the Logic of Scientific Discovery,* London: Hutchinson & Co.

Popper, Karl R., and Eccles, John C. 1977. *The Self and Its Brain: An Argument for Interactionism,* London: Routledge.

Quine, W. V. 1969. "Epistemology Naturalized." In *Ontological Relativity and Other Essays*. New York and London: Columbia University Press.

——— 1995. "Naturalism; or Living within One's Means." *Dialectica* **49**, 251-61.

Reichenbach, Hans 1971. *The Direction of Time*. Edited by Maria Reichenbach Berkeley and Los Angeles: University of California Press.

Robert Bosch GmbH 1993. *Automotive Handbook,* Stuttgart: VDI Verlag.

Rodd, Philip 1964. "Some Comments on Entropy and Information." *American Journal of Physics* **32**, 333-5.

Rosenberg, J.F. 1980. *One World and Our Knowledge of It,* Dordrecht: D. Reidel Publishing Company.

Russell, Bertrand 1959. *Mysticism and Logic and Other Essays,* London: George Allen & Unwin.

Ryle, Gilbert 1949. *The Concept of Mind,* London: Penguin Books.

Schlick, Moritz 1962. *Problems of Ethics,* New York: Dover Publications.

Schrödinger, Erwin 1967. *What Is Life?,* Cambridge: Cambridge University Press.

Schrödinger, Erwin, translated by John D. Trimmer, 1935/1980. "Die Gegenwärtige Situation in Der Quantenmechanik." *Proceedings of the American Philosophical Society* **124**, 323 - 38.

Schwartz, Jeffrey, and Begley, Sharon 2002. *The Mind and the Brain,* New York: Regan Books.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *Philosophical Review* **68**, 141-56.

——— 1963. "Materialism." *The Journal of Philosophy* **LX**, 651-62.

Smart, J.J.C. 1961. "Free-Will, Praise and Blame." *Mind* **LXX**, 291-306.

Squires, Euan J. 1988. "The Unique World of the Everett Version of Quantum Theory." *Foundations of Physics Letters* **1**, 13-20.

Stapp, Henry P. 1999a. "Attention, Intention, and Will in Quantum Physics." *Journal of Consciousness Studies* **6**, 143-64.

——— 2004. *Mind, Matter, and Quantum Mechanics,* Berlin, New York: Springer.

——— forthcoming a. "Quantum Approaches to Consciousness." In *Cambridge Handbook of Consciousness*, edited by P. Zelazo and M. Moscovitch. New York: Cambridge University Press.

——— forthcoming b. "Quantum Mechanical Theories of Consciousness." In *Blackwell Companion to Consciousness*, edited by Susan Schneider and Max Velmans.

Strawson, Peter 1962. "Freedom and Resentment." *Proceedings of the British Academy* **xlviii**, 1-25.

Szentágothai, J. 1975. "The 'Module-Concept' in Cerebral Cortex Architecture." *Brain Research* **95**, 475-96.

Szilard, L. 1964. "On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings." *Behavioral Science* **9**, 301-10.

Thomson, William 1852a. "On the Mechanical Action of Radiant Heat or Light: On the Power of Animated Creatures over Matter: On the Sources Available to Man for the Production of Mechanical Effect." *Proceedings of the Royal Society of Edinburgh*, [from Sir William Thomson, Mathematical and Physical Papers, vol. 1, pp. 505-10].

——— 1852b. "On a Universal Tendency in Nature to the Dissipation of Mechanical Energy." *Proceedings of the Royal Society of Edinburgh*, [from Sir William Thomson, Mathematical and Physical Papers, vol. 1, pp. 511-14].

Van Inwagen, Peter 1975. "The Incompatibility of Free Will and Determinism." *Philosophical Studies* **27**, 185-99.

——— 1983. *An Essay on Free Will,* Oxford: Clarendon Press.

——— 2002. "Free Will Remains a Mystery." In *The Oxford Handbook of Free Will*, edited by Robert Kane, 158-77. Oxford: Oxford University Press.

Victoria, Education Department 1952. *Playmates - the Victorian Reader, First Book,* Melbourne: Government Printer.

von Neumann, John 1955. *Mathematical Foundations of Quantum Mechanics*. Translated by Robert T. Beyer Princeton: Princeton University Press.

von Wright, Georg Henrik 1951. *An Essay in Modal Logic,* Amsterdam: North Holland Publishing Company.

Walter, Sven, and Heckmann, Hans-Dieter 2003. *Physicalism and Mental Causation,* Exeter: Imprint Academic.

Wang, G. M., Sevick, E. M., Mittag, Emil, Searles, Debra J., and Evans, Denis J. 2002. "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales." *Physical Review Letters* **89**, 050601.

Wegner, Daniel M. 2002. *The Illusion of Conscious Will,* Cambridge, Mass.: MIT Press.

——— 2004. "Précis of *the Illusion of Conscious Will*." *Behavioral and Brain Sciences* **27**, 649-92.

Weinberg, Steven 1994. *Dreams of a Final Theory*. 2nd ed New York: Vintage Books.

Whitehead, Alfred North 1929. *Process and Reality,* New York: The Social Science Book Store.

Wilson, David L. 1999. "Mind–Brain Interaction and Violation of Physical Laws." *Journal of Consciousness Studies* **6**, 185-200.

Wittgenstein, Ludwig 1953. *Philosophical Investigations,* Oxford: Basil Blackwell.

Wittgenstein, Ludwig, Anscombe, G. E. M., Wright, G. H. von, and Paul, Denis 1979. *On Certainty*. Reprinted with corrections and indices. ed Oxford: Blackwell.

Young, Hugh D. 1992. *University Physics*. 8th ed Reading, Mass.: Addison-Wesley.