Deliberation and Reason

Deliberation and Reason

Richard Baron

Matador
5 Weir Road
Kibworth
Leicester LE8 0LQ, UK
Tel: (+44) 116 2792299
Email: books@troubador.co.uk
Web: www.troubador.co.uk/matador

British Library Cataloguing in Publication Data.
A catalogue record for this book is available from the British Library.

ISBN 978 184876 250 3

Typeset in 11pt Sabon by Troubador Publishing Ltd, Leicester, UK
Printed in the UK by MPG Biddles, Kings Lynn, Norfolk

Matador is an imprint of Troubador Publishing Ltd

Contents

Preface

This book is about the thinking in which we engage when we reflectively decide what to do, and when we reflectively reach conclusions as to the correct answers to questions. Some philosophers rigorously separate our choices of action from our adoptions of belief, on the ground that we have choices as to what to do, but no choice as to what to believe. I treat the two together when considering processes of deliberation, but separately when considering the rationality of conclusions.

The main objective is to identify a way of looking at ourselves and at our deliberations that is adequate to our lives. It must accommodate both our conception of ourselves as free, rational and self-directed subjects, and the phenomenology of deliberation. It must also identify a place for us that will feel like home, doing justice to our status as subjects, within the world as we relate to it when we practise the natural sciences. The central claims are not about how we are, but about how we should look at ourselves. A key task is to show that this limited ambition, which is forced on us by the need to avoid metaphysical implausibility, nonetheless allows us to develop a position that has sufficient strength to do its work. The aim is to show something that is all too easily taken for granted. This is that we can limit ourselves to a strictly naturalistic ontology, while still having access to a generous idiom that allows us to speak of ourselves as free in the exercise of our rationality.

London is a wonderful city in which to engage in philosophy. I owe an enormous debt to the many people who run and participate in its seminars, allowing the all-important public exchanges of ideas that drive private thinking forward. I am also very grateful to the staff of the British Library, of the University of London's Senate House Library, and of Cambridge University Library. Finally, my thanks go to my publisher, and especially to the editor who saved me from many mistakes.

Richard Baron
Midsummer 2009

A note on references

References in the text give the author and the title. Full details can be found in the bibliography. I have given references in forms that are independent of edition, such as section numbers, when such references identify passages with sufficient precision. This approach should be increasingly useful with the rise of the electronic text. When page numbers are given, they relate to the editions that are cited in the bibliography. Roman numerals have been converted into Arabic numerals when they merely give volume, chapter or section numbers, rather than being parts of titles.

References to works by Plato and by Aristotle include Stephanus and Bekker numbers respectively. References to works by Descartes include volume and page numbers in the Adam and Tannery edition, in the form AT 7: 123. References to Kant's *Critique of Pure Reason* include the usual A and B page numbers, in the form A76/B101. References to other works by Kant include volume and page numbers in the Prussian Academy edition, in the form Ak.1: 234.

I have used English titles of works in other languages when the works are well-known by those English titles. In borderline cases, I have erred on the side of the original language.

English translations of texts in other languages are taken from the editions that are cited in the bibliography. When no English edition is cited, translations are my own.

CHAPTER 1

Deliberation

The principal topic of this book is how we deliberate. In the first three chapters, I concentrate on processes of deliberation. In chapters 4 and 5, the discussion broadens out to encompass both the conclusions that deliberators reach, and the relationship of those conclusions to the starting points of available information and of methods of using it. In chapter 6, I discuss some problems in the philosophy of science. That chapter ends with a link between the scientific view of the world and the proposals that are made in chapters 2 and 3.

Each of the first three chapters includes one stage in the discussion of the process of deliberation. In chapter 1, I set out the work to be done. I motivate the discussion, outline a model of deliberation, put forward two contrasting views of the process, the program view and the view that sees deliberation with mastery, and argue for the latter view. In chapter 2, I set out a way to see deliberation with mastery, argue that we must overlook the causal closure of the physical, and start to develop a picture of ourselves that will make the overall proposal acceptable. In chapter 3, I develop this picture further, argue that my proposal is both acceptable and adequate to the task, and consider other approaches that are in play in the same field.

In section 1.1, I start by distinguishing between processes of deliberation and the conclusions that deliberators reach, then set out why we can treat choices of action and adoptions of belief together when we are concerned with processes. I then set out the motivation for the argument. This is the need for a philosophy that is true to our lives, meaning both our

self-conception and our experience of deliberation. Finally, I discuss the idea of a way of looking at ourselves and at our deliberations.

In section 1.2, I set out a model of our processes of deliberation. A selection stage, in which we select both weights to attach to pieces of evidence and a method of argument, is followed by a calculation stage. I discuss the acceptability, and the limited implications, of the degree of idealization that is involved in the use of this model. I pick out the evidence of which the subject is aware, and the prior inclinations of the subject, as the rational antecedents of the process. I also address the risk of doxastic voluntarism.

In section 1.3, I set out the program view, the view that sees a process of deliberation as the execution of a program that is foisted on the subject. The contents of the program will reflect both the prior inclinations of the subject to think in certain ways, and the evidence that has come to the subject's notice.

In section 1.4, I first argue that it is important for us to attribute a strong form of freedom to ourselves. I then argue that the program view is inadequate to our self-conception as free, rational and self-directed subjects. I put forward a different view, that of deliberation with mastery. If we take that view, then we see the subject as freely making an intervention in the process that is extraneous to the program. I set out three arguments for taking that view, arguments that are respectively based on our self-conception, on our social relationships and on the phenomenology of deliberation. In section 1.5, I compare the view that sees us as deliberating with mastery with other views.

1.1 Our processes of deliberation

In this book, I argue for the desirability of taking a particular view of our processes of deliberation. I go on to set out what we need in order to be able to see our deliberations in that way, and the philosophical consequences of having what we need. We take the recommended view when we see ourselves as weighing the pros and cons in a certain way before deciding what actions to perform, and when we see ourselves as handling evidence

in a certain way before acquiring beliefs. Seeing a process of deliberation in the recommended way is, however, neither necessary nor sufficient for us to see the conclusions, the decisions or the beliefs, as rational. It is not necessary, because conclusions can be seen as rational so long as they are appropriately related to the information that was available. Our view of the process, abstracted from its content, is not then important. Seeing a process in the recommended way is not sufficient, because an eccentric can be seen as reasoning in that way even if he reaches a bizarre decision or belief. He may have a bizarre way of attaching significance to the pieces of information that are available to him, or a bizarre way of making use of that information.

We therefore need to distinguish between two ways of considering conclusions. If we take the recommended view of the process of deliberation, that supports our self-conception as free, rational and self-directed subjects. As a result, we respect the conclusions of one another's deliberations. We do not force people to act, or believe, differently, nor do we try to trick them into acting or believing differently. We confine any attempts to change their minds to rational persuasion. We are also inclined to take our opponents' views seriously even if we have confidence in our own, contrary, views. I shall attach the label "deliberation with mastery" to the view of the process that I recommend. Viewing a process of deliberation in that way not only supports an appropriate self-conception. It also accommodates the phenomenology of deliberation. I shall argue that to view a process in the recommended way, we must incorporate non-naturalistic elements, and that we must adopt a non-naturalistic conception of ourselves. We can also consider conclusions in another way, and ask whether actions or beliefs are rational. We then consider the content of the conclusions, and the content of the processes that gave rise to them. I shall refer to conclusions that are appropriately related to available information as rational actions or rational beliefs.

Actions and beliefs

I am concerned both with our choices of action and with our adoptions of belief in response to factual questions. Specifically, I am concerned with the choices and adoptions where systematic reflection is to be expected. We

choose many actions and acquire, or hold, many beliefs without reflection. There are some actions, such as walking to a nearby shop when one needs a bottle of milk, that it would be preposterous to choose by a process of systematic reflection. Likewise, it would usually make no sense for someone to reflect systematically in order to work out which town he was in, even when he was away on holiday. There are other actions that one could sensibly choose, and questions that one might sensibly answer, by a process of systematic reflection, although the effort would not normally be worthwhile. Examples include a choice between taking a bus and taking a train to the other side of town, and the question of how far the Moon might be from the Earth. (For the latter, one would just consult the most convenient encyclopedia.) Finally, there are decisions and questions where systematic reflection is not only natural, but expected. A decision on whether to invest a large sum of money in equities or in bonds would be one example. A controversial academic question, for example the question of which English king, Richard III or Henry VII, was responsible for the deaths of the Princes in the Tower, would be another. In the latter case, one would reflect on the evidence with a view to coming to believe one answer or the other, but without expecting to have a choice as to which belief to adopt.

It does make sense to treat actions and beliefs together. It is true that a process of reflection on evidence will ideally leave us with no option as to what to believe, whereas with actions we are not constrained in the same way. (We may be constrained in a different way, in that one action might strike us as the only sensible or moral action, but that would not be the same sort of constraint as the constraint to adopt beliefs that the evidence unambiguously favoured, unless the requirement to act in such a way as to achieve one's ends, or the requirement to act in accordance with a given system of morality, was some sort of fact about the world.) Despite this difference, there is an important similarity. The process of arriving at beliefs ideally involves systematic reflection on the available evidence, whenever the question to be answered is a sophisticated one, the truth is not obvious to the prospective believer, and it is not appropriate simply to defer to experts. Deference might be inappropriate because the experts were divided on the question, because the prospective believer was herself an expert, or because the question was of such a nature that no particular type of person

was especially authoritative. Likewise, certain actions are ideally chosen after systematic reflection on the pros and cons of the options. The similarity of the processes makes it appropriate to discuss deliberation in relation to actions and in relation to beliefs as a single type of process. When I set out a model of deliberation in section 1.2, it will become clear that the element of choice that is involved in relation to beliefs is unobjectionable because of its location in the process.

"Actions" and "beliefs" should be taken to refer exclusively to actions and beliefs that one would expect to be chosen or adopted following systematic reflection, except where otherwise specified. Choices of action and adoptions of belief in these categories make up only a modest proportion of human life, but they are still important. What we do in such cases is a measure of the human powers of thought that we rightly celebrate. How we see what we do can support our self-conception as free, rational and self-directed subjects.

Since my concern is the process of choosing actions or of adopting beliefs, rather than its outcome, the fact that the outcome may be more strongly constrained for beliefs than for actions will not affect the argument. It is worth noticing that the choice of an action and the adoption of a belief have in common, not only the process but its end-point, a commitment to the action or to the belief. That commitment may be provisional. Actions can be interrupted, and they may not even be started if there is a change of mind before the time for them. Likewise, beliefs can be changed. But there is still a commitment that marks the end of the process of deliberation. I shall not investigate the process of action beyond that moment of commitment. I shall, therefore, not be concerned with the links between the will and bodily mechanisms. Nor shall I be concerned with the puzzles that are set by phenomena like deviant causal chains, when an agent achieves an intended result by unintended means, his intended means failing or being pre-empted.

Motivation

The motivation for the argument is that we need a philosophy that is true to our lives, and one that finds places for the things that are important to us. This is not a need that we would feel if we were studying an alien race,

the members of which we did not consider to have mental lives that were anything like our own. A mere natural history would then be sufficient, and the standards of adequacy to be observed would be those of the natural sciences. But for ourselves, we need more. We would be discontented with an account of ourselves that was no more than natural history, an account that merely recited the facts as they might equally well appear to a non-human observer who had no sense of what it was like to live a human life. We would feel the pinch of Wittgenstein's remark, "In the world, everything is as it is and everything happens as it happens: there is no worth *in* it" (*Tractatus*, 6.41; on the interpretation of this remark and of the surrounding text, see Wiggins, "Wittgenstein on Ethics and the Riddle of Life").

I shall concentrate on two broad aspects of our lives, to which a satisfactory philosophy must be true. The first aspect is our self-conception. We see ourselves not as mere computers, but as free, rational and self-directed subjects who can stand back from evidence and decide how to proceed. The concept of the rational and self-directed human subject is under attack as "an invented Western tradition that has had its day" (Thrift, "I Just Don't Know What Got into Me: Where is the Subject?", page 84). But most of us still see ourselves as rational and self-directed subjects, and in so doing we honour a proud and productive tradition. The second aspect is our inner experience. I shall use the term "experience" in a broad sense, to include both the experience of external objects, such as trees, and the experience of states, such as the state of being in a room that is heated to 30 degrees centigrade or the state of having an empty stomach, whether or not the subject brings those experiences to reflective consciousness. I shall use the term "inner experience" to refer to our experience of having experiences, and also to our conscious experience of deliberation, of decision and of action. Inner experience is the species of experience that makes up a large part of the content of our inner lives. It includes both the phenomenology of the more basic level of experience of external objects and of states, and the phenomenology of deliberation, of decision and of action. The phenomenology of sensory experiences differs from the phenomenology of deliberation, of decision and of action, but deliberation, decision and action do have their distinctive

phenomenologies. If they did not, we could not compare our experiences of deliberation, of decision and of action, but in fact we can.

This does not mean that there is a phenomenology that is objective in the sense that a very wide range of rational beings, whether or not human, could grasp our phenomenology. My view is that there is no such objective phenomenology. But that does not preclude claims that the inner experiences of two creatures are in general similar, or that two specific events in the lives of different creatures are subjectively similar. Even if inner experiences cannot be compared directly in any objective fashion, objective biological similarity can make it very likely that the inner experiences of two creatures are in general similar. When they are in fact similar, the creatures concerned can recognize that similarity through their interactions, and in particular through their conversations, without having to invoke biological theory. Indeed, such recognition supports the claim that biological similarity is accompanied by experiential similarity. When general experiential similarity exists, it is also possible to pick out specific events and recognize their subjective similarity. This way of proceeding does not, however, give us any way of appraising two specific events in the lives of different creatures for experiential similarity, when we do not detect that the creatures have similar inner experiences in general. And there is no guarantee that biological difference would lead to experiential difference.

A consequence of the motivation, the need for a philosophy that is true to our lives, is that although the views that are put forward here might strike human beings as correct, they would not appeal to all rational beings. To the extent that a potential for universal appeal is a condition of objectivity, I cannot claim objectivity. As it happens, there is another reason for not claiming objectivity. The picture that I shall build up would not be plausible if it were claimed to state facts, rather than having its status limited to that of a legitimate view of the way in which we make some of our decisions and reach some of our factual conclusions. The legitimacy of the view does not imply that there is anything in the world, in addition to the things that could be mentioned in a mere natural history, or that those things actually have any non-natural properties. Despite these limitations, the resulting picture gives us worthwhile results, while not licensing the

free-for-all that might be feared to be the result of a move from telling it like it is to deciding how to look at our activities. We can get rich sounds from our blue guitars, on which we do not play things as they are (Stevens, "The Man with the Blue Guitar").

The notion of a view

I shall take the notion of a view, a way of looking at things, including ourselves and episodes in our lives, to be unproblematic. I shall not discuss this notion as a theme in itself. The success of the argument in section 3.5, to the effect that something as weak as a way of looking at things can do the work that needs to be done, will be essential to the acceptability of this relaxed attitude. That argument needs to show that nothing stronger than a mere way of looking at things is required. Anything stronger might well be philosophically problematic.

Others would not be so relaxed. Jürgen Habermas, for example, attaches great significance to the epistemic dualism that comes with the distinction between the impersonal observer perspective and the social participant perspective. He considers this dualism to be unavoidable because of the interlocking of the perspectives. That interlocking is itself unavoidable because scientific observers are inevitably engaged in social scientific practice ("The Language Game of Responsible Agency and the Problem of Free Will"). For Habermas, the challenge is to reconcile the two perspectives within a single overall picture. If one sets up the problem in the terms that Habermas chooses, that is a formidable challenge. If I were to cast the view of our processes of deliberation that would be given in a natural history and the view that I recommend in the respective roles of the impersonal observer perspective and the participant perspective, admittedly an imperfect casting, I would be faced with a comparable challenge. But I do not need to set up the problem in the terms that Habermas chooses. Habermas is disinclined to allow the self-sufficient existence of scientific knowledge, in that he will not allow the participant perspective to be subordinated (ibid., pages 30-31). He also has a strong sense of the semantically mediated causality of reasons (ibid., pages 29-30). I differ from Habermas on the first point. On the second point, I do not rely on the existence of any causality beyond physical causality. These two

differences allow me to see the interlocking of perspectives as looser than he would allow. Consequently, I do not find the interlocking to be as problematic as he considers it to be.

Having said that, I do seek a complete picture that is, so far as possible, integrated. I am not content merely to introduce a convenient way of looking at ourselves or at episodes in our lives. I therefore set out the relationship between the view that I recommend and the view of ourselves and of the world that is given by the natural sciences. My desire to integrate our subjective view of our deliberations with our view of the world as a collection of natural objects that interact with one another causally is one of the driving forces of the argument. That desire leads me to recommend seeing extraneous interventions in the causal network, interventions that give a direct link between the two views, and then to analyse the supposed origins and operation of those supposed interventions to the greatest possible extent. Sadly, this turns out to be far less than the extent to which the origins and operation of ordinary physical causes can be analysed. I also discuss the element of mystery that attaches to my proposal by virtue of the narrow limits of the analysis. I recognize the need to argue that this element of mystery is acceptable.

Finally, the view of our processes of deliberation that I propose is a view that we do not often adopt explicitly. Rather, it is implicit in our self-conception and in our attitudes toward others. It is most likely to come out into the open when we consider how to accommodate the phenomenology of deliberation. But even a view that is mostly implicit in our ways of thought needs to be examined for its implications and for its fit with our other thoughts. In what follows, I shall generally take the implicit nature of the view as read, but I shall issue occasional reminders.

1.2 A model of deliberation

In order to discuss our processes of deliberation, we need a model of them. I shall set out a model in some detail here. This will help to make clear the work that needs to be done.

Suppose that an agent has to choose between a range of options, and

that the choice is one of sufficient significance and complexity that it is to be made through an articulated process of reasoning. She will select a method of argument to use. She will also consider pieces of evidence, taking some of them to be weighty enough that they should have considerable influence over the choice. (I shall use the term "evidence" to cover information that can influence a decision as to what to do, as well as information that can lead us to adopt beliefs. This will include information about the agent's current desires.) Then she will use the chosen method of argument in order to move from the evidence to a decision. How she chooses to argue, and how she weights the evidence, will both be consequences of two things. The first is the evidence that has come to her notice, some of which she may have sought out following preliminary deliberation on what sorts of consideration should bear on the decision, and some of which may influence how other items of evidence are weighted. The second is her set of prior inclinations. Those inclinations will encourage her to reason in certain ways, and to regard certain types of evidence as more significant than other types. The prior inclinations may result from prior factual knowledge, from prejudices, from character traits, from specific mental capacities or their absence, or from any other feature of her mind. She may be conscious of some inclinations, for example, a policy of reasoning as mathematically as possible, and unconscious of others, for example, an aversion to doing anything that might lead others to criticize her. The latter inclination might lead her to attach great weight to evidence about other people's opinions, without her realizing why she did so.

Consider, for example, someone who is wondering whether to leave her job in banking and re-train as a social worker. Evidence will include levels of earnings and of job security in banking and in social work, the success rate in the social work training programme, her feelings as to whether she enjoys work in banking and whether she would enjoy social work, and her views as to how work in each of the occupations would fit with her values, such as a value of promoting economic growth or a value of helping people who are in difficulty. She will weight these pieces of evidence, perhaps attaching great importance to her own financial security or attaching great importance to the degree of alignment of each

occupation with her values. She will choose a method, for example to follow the recommendation of the single most heavily weighted consideration or to accommodate the greatest possible number of reasonably weighty considerations. She will then argue to a conclusion, which will be her substantive decision.

The process of arriving at a belief will parallel this. The subject will have a question, and will wish to answer it. He will select a method of argument, assemble evidence, weight the evidence and then reason to a conclusion. Within the natural sciences, there may be little rational choice as to method, or as to weights to attach to pieces of evidence. One recalcitrant piece of evidence can be enough to bring down a whole theory. But even in the natural sciences, there are still important choices to be made. It is just that the choice of method may be a choice of the type of experiment, a choice that must be made in order to allow the gathering of evidence. And even if there are no choices to be made, the model can still be applied in a degenerate form. In the humanities, we can expect more choice. A historian of ideas, for example, who wanted to discover the lines of influence that related a previously neglected author's work to the works of earlier and later authors, might decide that the appropriate method was to establish the views of the author and of those who came before and after him on a range of significant questions, and then to draw up family trees of those views. He would then go out and review the texts to find the necessary views. Or he might heed Quentin Skinner's warning against the mythology of doctrines, the idea that each author must have had a view on each one of a given set of topics (*Visions of Politics: Volume 1, Regarding Method*, chapter 4, section 2). Then he would adopt a different method, and collect different evidence. As to the weighting of evidence, a medievalist who wanted to know what really happened on a given occasion but who was dependent on limited sources that omitted details, or that sometimes contradicted one another, might decide to attach great weight to the accounts that were given by one or two chroniclers, because they were known to have given accurate accounts of other events. Alternatively, he might select as method a careful analysis of all of the available sources, noting their agreements, their contradictions, how they came to be created and the implications for their reliability, and let the weighting of evidence

emerge from the application of that method. (For an example of this latter method, see Mortimer, "The Death of Edward II in Berkeley Castle".)

An actual process of deliberation may be both more complex and more haphazard than the model would suggest. For choices of action, it is almost certain to be more haphazard. But I do need to set out a clear and well-defined process of deliberation, in order to conduct a philosophical argument. While the model may be empirically inadequate to actual processes of deliberation, idealization can be justified, given my purpose. This is to ground our self-conception. The best ground is likely to be our supposed use of an ideal process of deliberation. If a choice of action or an adoption of belief is seen as the outcome of some well-defined process, that will greatly facilitate our regarding it as reached in a way of which we can be proud, rather than merely as reached. So long as my idealization, in the sense of simplification, is also an ideal, in the sense of something admirable, and so long as it is also not greatly at variance with reality, idealization should do no harm to the argument. The acceptability or otherwise of a given level of divergence from reality need not be measured purely by seeing how many steps in the idealized process can be matched with steps in actual processes. Elizabeth Anscombe noted, in relation to practical reasoning, that it sets out steps that can be revealed by analysis: "But if Aristotle's account were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order which is there whenever actions are done with intentions" (*Intention*, section 42). Likewise, my idealization can do its work if it reflects a pattern that is there to be uncovered.

There is one obvious respect in which the idealization is at variance with reality. This is the distinction between weighting pieces of evidence and selecting a method of argument. These cannot in practice be separated. A method of argument may well incorporate an approach to weighting evidence, and the attachment of particular significance to certain types of evidence may recommend one method of argument over another. There are also instances in which the selection of a method is absorbed into the stage of weighting, by virtue of the content of concepts. If, for example, there is evidence that a given action would significantly benefit a close friend, and that evidence is given great weight, that must count very

strongly in favour of performing the action. Only methods of argument that respected the content of the concepts of friendship and of benefit could sensibly be selected. If there were no other weighty considerations, nothing would need to be added to the method beyond such respect for the content of the concepts. A decision to perform the action would follow directly. But such considerations will not matter, because I shall treat the weighting of evidence and the selection of a method of argument together. They are identified separately in order to show what is involved, not to suggest that they must be undertaken separately.

More generally, we must be careful not to read too much into the idealized picture, because doing so might give rise to the misleading appearance of more serious divergence from reality. Two things in particular should be noted. First, there is no implication that we engage in calculation that is free of prejudice. Prior inclinations influence selections, both of weights and of methods. Second, the idealization does not entail that the process be one of collecting premises for a deductive argument, and then reasoning after the manner of the examples in a textbook of logic. We are concerned with arguments that have starting points, in the form of pieces of evidence, but the arguments can be in any style. My idealization is, for example, compatible with the models-based form of reasoning that is described by Philip Johnson-Laird in *How We Reason*. In his view, we reason by constructing mental models in ways that allow us to simplify the reasoning process. We may, for example, reason deductively, with a risk of error, by constructing some models of the world that take into account some of the pieces of evidence (ibid., chapter 8). If we were to take the modelling process to be something that we chose, then a decision to use a limited number of models that reflected limited parts of the evidence could be regarded as the selection of a method. The decision as to which pieces of evidence to ensure were taken into account by the models could be regarded as the selection of weights to attach to pieces of evidence. If a piece of evidence was treated as one that had to be taken into account, then it would ipso facto be regarded as reasonably weighty. The selection of which models to construct might also be regarded as the selection of weights. We would construct models in which the pieces of evidence that we considered to be weightiest played significant roles. If, on the other

hand, we inevitably reasoned by building models, my model of deliberation could still be applied. The place in the model-building process of the selection of weights could survive as just described. We could also still find room for a choice of method, in that we could use models as ways of carrying through deductive, inductive or abductive reasoning. Deduction would require the construction of a set of models in which pieces of evidence were directly represented as propositions that were undoubtedly true. Induction would require the construction of models that incorporated various presumed regularities in the phenomena that were of interest, and that also incorporated relevant observed regularities. The presumed regularities could then be tested to see whether they could legitimately be inferred from the observed regularities, using mathematical techniques such as correlation and the analysis of variance. Abduction would require the construction of models that incorporated various supposed facts in addition to the evidence, and then the selection of the model that gave the most satisfying picture overall.

We can identify the evidence that has come to the notice of the subject, and the subject's prior inclinations, as the rational antecedents. The word "rational" does not in this context indicate rationality in the evaluative sense. It merely indicates that these antecedents exist in logical space, taken to be the home of propositions as well as of concepts, rather than in the physical world. Thus the prior inclinations should be taken to be propositions, such as the proposition that the subject is averse to risk, rather than their being facts that could be reported in propositions. Likewise, pieces of evidence should be taken to be propositions. (The fact that the subject will actually be driven forward by physical realities, and not by propositions, does not matter. We identify the rational antecedents in order to understand the logical course of the subject's deliberations, so it is appropriate for us to think in terms of propositions, and it does not matter that they are causally inert.) The rational antecedents can be seen as collectively setting the subject on a course that leads to a decision or to a belief. They lead to the selection of weights to assign to pieces of evidence, and to the selection of a method of argument that will use those weighted pieces of evidence. I shall call this the selection stage. It is followed by the calculation stage, in which the selected method is applied

to the weighted evidence. The final result is the substantive decision as to what the subject is to do, or the adoption of a substantive belief as to the correct answer to the question posed. The selections of weights and of a method are merely intermediate decisions.

We can now see how objectionable doxastic voluntarism is avoided, in cases where the subject deliberates in order to answer a question and thereby add to his stock of beliefs. The idea that we might choose our beliefs has been opposed by David Hume (*An Enquiry Concerning Human Understanding*, section 5, part 2) and by Bernard Williams ("Deciding to believe"), among others. Not everyone accepts the case against voluntarism, but the case is strong enough that voluntarism is best avoided. The element of choice that is implicit in my model of deliberation is located too early in the process for it to fall to the standard objections to doxastic voluntarism. It is merely a choice as to how to weight the evidence and how to work from the evidence to a conclusion that the subject must then accept, unless the subject regards the conclusion as so implausible that he goes back and enquires as to whether he made the right choices at the selection stage. This is not to allow that a subject can choose a belief and then find a way to get to it by making appropriate choices at the selection stage. Such manipulation is ruled out by our understanding of what it is to investigate a question with a view to finding out the answer. That is, it is ruled out by an extension of Williams's argument against doxastic voluntarism, that it is characteristic of beliefs that they aim at truth ("Deciding to believe", page 148). It is also worth noting that a belief that someone holds following systematic deliberation can hardly be held by him without deliberate affirmation. It would, therefore, not be appropriate to identify an independent voluntary companion to such a belief, whether acceptance (Cohen, *An Essay on Belief and Acceptance*) or opinion (Dennett, *Brainstorms*, chapter 16).

1.3 The program view

If we see an instance of deliberation as following a course that is indicated by the rational antecedents, then we can see it as the execution of a

program. The content of the program would be determined by all of the rational antecedents, both the prior inclinations and the evidence. Its execution would include both the selection stage and the calculation stage. There is no need to see the program as actually encoded in the subject's head. Setting out the program would merely amount to describing a pattern that was there to be uncovered. What is important is that if we see the course of a deliberation as the execution of a program, and if we are aware of the content of the supposed program, then we see the way in which the deliberation must go, or the range of courses that it may follow, if the steps taken are all to accord with some logic of reasoning. In so doing we extract all that can be extracted from the rational antecedents by way of explanation of the course of the deliberation. The taking of any course that fell outside the range would not be explicable merely by reference to the rational antecedents. I shall use the term "the program view" for the view of deliberations that sees them merely as the execution of programs, the contents of which are derived from the rational antecedents, without any interventions that are extraneous to the programs. Note that because each program captures the entire contents of the rational antecedents, including not just prior inclinations but also current evidence, the program will change as new information is acquired. Similarly, the programs that were associated with different deliberations by the same subject might be significantly different, although stability of character would imply stability of prior inclinations, so that there would be similarities between the programs.

If a deliberation is regarded in that way, the subject will be said to follow the program. I shall also refer to the making of choices that could be made on the execution of a supposed program as complying with the program, and to the making of other choices as going outside the program. There is a vital distinction between following a program and complying with it. When a subject is seen as following a program, its execution is the only thing that is seen as directing her, the only explanation of how she decides. It may not direct her every choice, but choices that are not explained by the program are not explained at all. The fact that she complies does not, however, mean that the program's execution need be seen as the only source of direction. We might also see an extraneous

intervention. But for following not to be seen when there is compliance, we must see an intervention that made no difference to the course of the deliberation, but that would have made a difference if it had been different, and that would have explained the difference made. It is not enough to see the absence of an intervention that could have been present. This point will come up again in section 1.4, when I propose seeing extraneous interventions, and not merely the possibility of such interventions.

A program cannot in the abstract direct anyone. Only its execution, which imports the notion of pressure to move onward in accordance with the instructions in the program, can do that. But we can think in terms of action of the mental on the mental, if we wish. Direction solely by the program would at the mental level amount to the subject's having no extraneous thoughts, but only the thoughts that one would have in executing the program. Physicalism in the philosophy of mind is not required in order to make sense of this notion of direction, although it is not excluded either.

The rational antecedents will often not be sufficient to allow us to see a program that could only follow one course, even if it were properly executed. But that will not matter. The program would still capture all of the explanation of the course of the deliberation that could be extracted from the rational antecedents. It is also important not to see the program as consciously written or chosen by the subject. It should be seen as foisted on the subject, because the subject does not at the highest level choose how to think. Choices that the subject makes as to how to proceed are made within the selection stage, which can be seen as the execution of the first part of the program. Those choices may include the conscious construction or selection of lower-level programs that the subject then uses in order to carry through the deliberation, including programs that help the subject to make choices that arise within the context of the deliberation as a whole.

We can only identify courses for a program to follow if we use some logic or other. Should not that logic too be regarded as something that is chosen at the selection stage? And would not its being so regarded undermine the very idea that there was any set course for a program to follow, in advance of its execution? There is something in this objection, just as there is something in Quine's view that even the most central

elements in our fabric of belief, our logical laws, could in extreme circumstances be candidates for revision ("Two Dogmas of Empiricism", section 6, pages 42-43). But we need not be overly concerned. In order to have any discussion at all about deliberation, we must assume that some sort of logic should be regarded as given, and not as chosen on each occasion of deliberation, even chosen implicitly or by default. The logic that is taken to be given should not be a logic that we would regard as specific to a given type of human being, or even as specific to human beings generally, as opposed to other rational beings. If it were such a specific logic, its use would be better classified as following from rational antecedents at the selection stage, and therefore as selected in the course of execution of the program. (It could not, however, be seen as following from rational antecedents unless at least some minimal logic were already in place.) We should also be wary of parts of logic, the legitimacy of which is seriously disputed. We might, for example, feel safer if we could manage without parts of logic to which intuitionists would object. On the other hand, there would be no reason to exclude parts of logic merely because there were arguments for treating them as parts of mathematics rather than as parts of logic itself. And once worries about contested parts of logic had been laid to rest, the topic-neutrality of logic, together with its close integration with mathematics, a body of knowledge that gives every indication of being the key to the structure of the whole of the Universe, would reassure us that we could legitimately regard a substantial part of standard logic as given. That would be enough to yield a course, or a limited range of courses, that a process of deliberation should follow, given the rational antecedents. That in turn would support taking the program view.

The outcome of the execution of a supposed program, the substantive decision or belief, might not be inevitable, even if deliberation accorded with the given logic. Random elements might be present in the program, either at the selection stage or at the calculation stage. The subject might, for example, have cast a die at some point, perhaps because her rational antecedents led her to want the thrill of trusting to chance. Leaving such thrill-seeking aside, there are many occasions on which the rational antecedents are insufficient to yield unique selections of weights and of

method, or on which the application of a given method of argument to the available evidence is not enough to calculate a unique outcome. The subject just has to decide, for no reason that could be given within the subject's own reasoning. Such an occasion can also be regarded as one on which a die is cast. The die may, however, be loaded, reflecting prior inclinations or the available evidence, so that one outcome is more probable than others. Indeterminacy at the level of reasoning can also arise out of the subject's use of concepts that are not precisely defined. I shall not treat mere indeterminacy, from whatever source it may arise, as taking us beyond a process that can be regarded as the execution of a program, whether or not we can identify statistical regularities, for example, that a given outcome arises 70 per cent of the time. This does not, however, mean that we should routinely expect substantial under-determination. Sometimes only one outcome will be at all likely. If an action is being chosen, there may be only one obvious thing for someone of a given character, as represented by her prior inclinations, to do. If a question is to be answered, the evidence may overwhelmingly favour just one possible conclusion. And in any case, we should not regard a lack of determinants of the course of a deliberation as leaving open a choice of programs, because the subject is not to be seen as choosing the program at all. There will be one program that will incorporate all and only the information that can be extracted from the rational antecedents. The effects of a lack of determinants should be reflected within that program, by building in appropriate indeterminacy.

The selection stage is not often conspicuous to us. It is easy not to see it as the execution of the first part of a program, because it is easy not to see it at all. It often involves no conscious reflection. Some evidence just strikes us as particularly significant, and we just start to argue from the evidence to a decision or to the adoption of a belief, without being aware of selecting a method of argument. Sometimes a subject does stop and think about her selections of weights and of method. But in all cases, we can see the process of making selections as moving like the execution of a program that is generated by the available evidence and the subject's prior inclinations, and that has as outputs the selections that are made. The process may amount to a movement to only one possible result that is explained by prior inclinations alone, with current evidence having no effect. Examples are

when the subject only ever uses one method of argument for decisions of a given type, such as a minimax method, and when she always attaches the greatest weight to evidence of a particular type, such as evidence as to how the well-being of her family would be affected. Even then, the process is there, and it can be seen as the execution of a program.

The process has so far been described at the level of reasoning, not at the physical level of brain cells. Alongside the evidence and the prior inclinations, we have states of the subject's brain that correspond to awareness of the evidence and to inclinations to think in certain ways. I shall identify these states of the brain as the natural antecedents. They participate in the causal mechanism. They underpin the mechanical nature of deliberation, making it possible for the subject to engage in extended reasoning. The fact that a causal physical process and a process of deliberation that operates according to rules of logic should march in step need not be a great mystery. In order to explain the fact that this happens, we need evidence of a physical organ of sufficient complexity to allow its states and changes to encode processes of reasoning, along with good reason to expect the processes of reasoning to be correlated with the states and changes of the physical organ. It is obvious that we have organs of sufficient complexity, although the details of how our brains work are not at all obvious. We should also expect correlations, because we have devised processes of reasoning that fit the capacities of our brains to handle them. To quote Angela Carter, "Our external symbols must always express the life within us with absolute precision; how could they do otherwise, since that life has generated them?" (*The Passion of New Eve*, page 6). A full explanation would, however, require more, because we are to a significant extent constrained by external facts to use some methods of reasoning rather than others. We cannot simply select whatever methods of reasoning happen to suit our grey matter and expect to be successful. We must use sound deductive systems, because we live in a world that stands as a model to our premises. We therefore need to explain how it is that we have brains that can use the methods that work, when our brains might have been just as complex but unable to use those methods. That is a question for evolutionary biology, rather than for philosophy. I return to the coincidence problem more generally in section 4.2.

1.4 The case for mastery

We wish to support our self-conception as free, rational and self-directed subjects. The program view is perfectly adequate on the side of rationality. It relates the course of a deliberation to the rational antecedents, making that course explicable. The problem lies on the side of freedom and self-directedness. I can now indicate the way in which the program view would be inadequate, and how to remedy that inadequacy. Under the program view, the subject is seen as following a program, without making extraneous interventions, and not merely as complying with a program. The lack of extraneous interventions will be central to the inadequacy. The remedy will consist in seeing such interventions.

We must first consider the reasons why we should see ourselves as free. The efforts of philosophers to accommodate freedom of the will in a causally determined world make it clear that freedom matters to us. But there is a more specific reason for wanting freedom, which can persuade us that the freedom that we need to see is of a sort that is incompatible with determinism. I submit that the primary need is not to support our self-conception as free, but to support our self-conception as self-directed. We want to feel that we, as subjects, are in charge of ourselves, that we lead our lives, and specifically our deliberations, and that we are not led. We matter, as subjects, before we even get close to moral considerations. Moreover, our conception of ourselves as subjects who are in charge must be significant. It must add something to the overall picture.

This is why the program view is inadequate. That view allows us to look at the subject in either one of two ways, and each is as good as the other. On the one hand, we can see the subject as in charge. The program that she follows is her program, and it reflects her characteristics and knowledge. She has guidance control (Fischer, "Compatibilism", section 8). On the other hand, we can see the program that is foisted on the subject as in charge. The subject then appears merely as an obedient part of the mechanism of the world. Nothing is gained or lost when we switch from one way of looking at the subject to the other. The accounts that are given are equivalent. We could even stop talking in terms of subjects, as opposed to biologically delineated parts of the mechanism. The subject could

dissolve into the mechanism. She would then be seen simply as driven by the physical mechanism, although we could still describe the course of her life in the language of programs, by talking about her evidence, her inclinations and her reasons. So under the program view, nothing is gained by identifying the subject as in charge, and the subject is liable to be demoted. The subject is, therefore, not significant enough to answer to our self-conception.

If, on the other hand, we see ourselves as having a freedom that is incompatible both with determinism and with the program view, we can secure the importance of the subject. We are to see ourselves as having alternative possibilities, as being able to do other than we in fact do. Specifically, we are to see ourselves as able to go outside our programs, and to do so under our control rather than by accident. That would be incompatible with the program view, because choices that were seen as possible but that would not be explained by a subject's program would still have explanations. They would be explained as the subject's choices. We can also secure the indispensability of the subject, and of the subject's position in charge. The subject would be indispensable because the possible choices would need to be seen both as controlled and as explicable, and not as accidents, so the subject would have to play her explanatory role. And switching to the view that subordinated the subject to the program would mean losing something, because it would not then be possible to see explanations for choices that were not explained by the program. Finally, seeing us as having the proposed type of freedom would be incompatible with seeing us as determined. I shall explore the relationship with determinism in section 2.2.

This, then, is the challenge. If we can find a way to see ourselves as free in this strong sense, a sense that certainly measures up to our self-conception, then we can also see ourselves as self-directed in a way that measures up to our self-conception. I shall in due course support this argument for seeing ourselves as free with other arguments for seeing freedom. Some of those other arguments will not be so hostile to compatibilist conceptions of freedom, but they will be perfectly accommodating to incompatibilist conceptions. There may be other ways to secure an appropriate sense of self-directedness. But the fact that

incompatibilist freedom can do the job so straightforwardly, together with the existence of the other arguments for seeing freedom, means that it is worth concentrating on this line. It is also worth noting that if we see a free choice between a range of alternatives, that gives content to the notion of being self-directed. If we saw someone as self-directed but as limited to directing himself along one course, the self-direction would seem rather hollow.

Having set out the link between seeing freedom and seeing self-directedness, I shall concentrate on freedom. But first, two asides are in order. The first one is that although we are to see ourselves as having the strong freedom that has been described, we are not to see ourselves as actually exercising it to the extent of doing something that is not naturalistically explicable. The second aside is that we should not think of avoiding the argument that we need to go beyond the program view by thinking of a subject as choosing a program. The program that is seen should be the one program that captures all of the information that can be extracted from the rational antecedents, and that does not add anything more. If we were to seek freedom under the program view by seeing the subject as able to choose one from a range of different programs to run, we would be brought no nearer to a picture that measured up to our self-conception. We would need to see the choice as made by following a super-program, in which case we would see that as the one comprehensive program, or as a choice with no identifiable source, or as a choice that was equivalent to a choice to go outside a program.

If we are to see ourselves as free in the required way (henceforth simply "free"), we must see any given subject as able to make choices of sets of weights and of a method that go outside the supposed program, without seeing her as cast adrift. We must see that any such choice would be made under the subject's control, and that it would have an identifiable source in the subject rather than coming from nowhere. The source must be a deliberate intervention that is extraneous to the program. Furthermore, it would not be enough to see the subject's control over the course of a deliberation merely as a control that could hypothetically have been exercised if the subject had gone outside the program, but that was not exercised because the subject in fact complied with the program. When the

subject makes an actual choice of weights or of method that complies with the program, we must see her as exercising the same type of control that we would have seen as exercised if she had gone outside the program. The extraneous intervention must be seen as made, even when it would make no difference. There are two reasons for this. The first is that our self-conception involves seeing ourselves as actually exercising our freedom, not merely as holding the ability to do so in reserve. The second is that the phenomenology of free and controlled choice arises in actual deliberations. This need, to see the subject as actually proceeding in the same way that we would want to see her as proceeding if she were to go outside the program, highlights a significance of our taking it that we have free choice. It is one that is just as important as the usual significance of giving us a satisfying feeling that we could, if we wanted, jump out of the causal tramlines that steer the world's development. By considering the personal control that we would want to see as making up for the lack of worldly controls with which we would be faced if we did jump out of the tramlines, we can see the personal control that we expect to exercise even when we remain within those tramlines.

What this amounts to is that we should see the subject as exercising mastery over the selection stage. The subject should be seen as exercising control over the actual selections of weights and of method, in the same way that she would be seen as exercising control if she were to go outside the program. She must be seen as actually making extraneous interventions. Then she will not be seen as following the program, but as choosing how to proceed, even if her choice actually amounts to compliance with the program. That is what I mean by deliberation with mastery. (She need not be seen as considering whether to comply with the program, as a question in which the program is mentioned. She need only be seen as using her freedom. Nor need she identify individual rational antecedents as such, although she must identify pieces of evidence in their role as evidence.) I shall now fill out this picture a little, before supplementing the argument for seeing ourselves as deliberating with mastery that has just been given, the argument that is based on our self-conception, and adding two more arguments. I shall discuss the implications of seeing mastery in chapter 2.

There are usually many logically possible options at the selection

stage, even if, on account of physical determinism, only one set of weights and only one method are actually available to a given subject on a given occasion, and even if complying with the program would make only one outcome of the selection stage possible. (I discuss the link between physical determinism and compliance with the program in section 2.2.) We do not need to see all of the many logically possible options as open to the subject. But we should see the subject as having a free choice between all of the options that she could consider, given her general mental capacities and her background knowledge. When I write of the options being seen as open, or of the generally available options, or of the possible sets of weights or the possible methods, I shall mean this range, narrower than the full range of logically possible options, but wider than the range that immediate circumstances and physical determinism would allow. The range is also wider than that which the subject's prior inclinations would allow. That is, she should be seen as well able to break free from her specific habits of thought. They are not to be assimilated to the general mental capacities that are to be seen as limiting the subject. The distinction is in general clear enough, even though there are borderline cases in which something may be treated either as a specific habit of thought, or as a limit on general capacities.

It is the selection stage that matters, because once selections have been made, a subject enters the calculation stage and there is only one rational way of carrying that through, although that way may involve deliberately inserted randomness or the use of ill-defined concepts, so that the outcome may not be inevitable. A re-consideration of where the calculation was going while it was still in progress could be perfectly rational. That would, however, be a re-consideration of the selections of weights and of method, rather than being either an arbitrary decision from within the process of calculation to change course, or the commission of an error in calculation. Likewise, a discovery in the course of the calculation stage that there was not enough material to lead to a unique outcome could quite reasonably lead to a decision to cast a die. But that would amount to re-opening the selection stage, in order to choose a method of argument that involved casting a die whenever the calculation stage would otherwise grind to a halt.

In outline, there are three arguments that favour seeing mastery over the selection stage. The first argument is that doing so supports our self-conception as free, rational and self-directed subjects. We see ourselves as standing back and deciding freely how to conduct our deliberations, where the available choices include choices that would mean over-riding the influence of rational antecedents and thereby going outside our supposed programs. Our rationality then consists in our making sensible choices, and in implementing those choices properly in chains of reasoning that are sometimes elaborate. The second argument is that seeing open choices as to how to proceed has the highly desirable effect of giving us good grounds to debate with other people properly, considering their views on merit and not dismissing what they say because we disapprove of their known or assumed ways of thinking. That desirable attitude toward people helps to make us members of a society, rather than members of a pack. The third argument is that seeing open choices accommodates the phenomenology of deliberation.

The argument from self-conception

The basic argument from self-conception has already been given. We need to see ourselves as having a full range of options, including options that would involve going outside programs, and as selecting whichever option we do take in a way that keeps us in control. That argument can most easily be supplemented in relation to adoptions of belief.

If we see ourselves merely as following programs, that does allow us to view the adoption of beliefs as a process that takes account of evidence, and that uses respectable methods of argument from evidence to conclusions. But we do not regard our beliefs merely as foisted on us by pieces of evidence. (Remember that I am specifically concerned with the formation of beliefs that are answers to questions where there is no substitute for reflection, and no easy recourse to experts who already know the answers. These are not cases in which the significance of the evidence is obvious.) We regard ourselves as choosing to attach significant weight to certain pieces of evidence, and as choosing how to use the weighted evidence to argue to a conclusion. Of course we have our habits of thought, which steer such choices, and we may be very glad to have habits that we see as particularly effective in steering us toward the adoption of true

beliefs. But we see ourselves as free to start approaching the world differently if we wish. We would see such a change of approach as a free choice, to which the subject was not led by antecedents but which was also not a random or uncontrolled change of mind. The independent mind that we admire is to be seen as potentially independent of its own inheritance, as well as being independent of the thoughts of others.

If someone does not see himself or herself as deliberating with mastery before adopting at least some beliefs, no knock-down argument is available. If anyone does not think that it is important to see himself or herself as someone who adopts beliefs in the light of reflection, but not as someone who arrives at beliefs mechanically, my advice is to look within and to consider what one expects of oneself as a human being. Two pointers may help.

The first pointer is based on what we expect of other people. We assume that we can have rational discussions with others, and that we can persuade them, or they us, by putting forward reasons for adopting various beliefs. (One analysis of what we expect can be found in the discussion of the conversational stance in Pettit and Smith, "Freedom in Belief and Desire".) If we thought it appropriate to see other people as acquiring beliefs by following programs, so that their beliefs or methods of thinking could be changed by putting in their way whatever stimulating propaganda would do the trick, we would not have the esteem for them that we do have. We would lack that esteem, even if we actually had no inclination to use propaganda or other tricks. This would be so, even if we saw other people as following sophisticated programs that had layered structures with higher-order rules of procedure to control the application of lower-order rules, that made thorough checks against existing beliefs and against standards of evidence before new beliefs were acquired, and that had been developed through experience and through internal reflection. If we would have a low regard for people whom we saw as acquiring beliefs by following programs, rather than by standing back and deliberating in ways that were not predetermined by antecedents, then we cannot rest content with the same conception of ourselves.

This consideration incidentally makes clear the strength of the conception of mastery that is involved. Someone who is seen as acquiring

beliefs by following a sophisticated program with a layered structure is still seen as open to manipulation. Sophistication would make it possible to use stimulating propaganda to change people's inclinations to reason in certain ways, or their inclinations to weight pieces of evidence in particular ways, rather than to change beliefs directly. Consider the rational animal, possessed of selective attention and a capacity to deliberate, whom Martha Nussbaum sets at the centre of Aristotle's theory of human development (*The Fragility of Goodness*, pages 285-287). Admirable though that creature is, he could still be seen as acquiring beliefs by following a sophisticated program, unless we went beyond Aristotle and saw him as deliberating with mastery.

The second pointer is that seeing ourselves as beings who adopt beliefs from a position of mastery, and not as beings who arrive at beliefs by following sophisticated programs, is satisfyingly incompatible with seeing ourselves as mere links in the causal chains of the world. Not only would few of us like to think of ourselves as mere links. Our experience of living indicates to us that we are not mere links. Those who maintain that there is no valid characterization of the world but the scientific characterization, could argue that the experience of living was deceptive. They could point out, correctly, that we may soon have an explanation, in neurological terms, of how we get the feeling of being more than mere causal links. The existence of such an explanation might be thought to lead to the immediate victory of those who considered only the scientific characterization to be valid. But that would not follow. We could acknowledge the scientific result in theory, and it would change our conception of ourselves just as profoundly as the theory of evolution has changed our conception of ourselves. But for the purposes of living, we would still see ourselves as more than followers of programs.

The supplement to the basic argument from self-conception in relation to the choice of actions follows the same pattern as the supplement in relation to the adoption of beliefs. Often we act out of habit, or because of momentary impulses which, on reflection, we may regard as irrational or at best non-rational. But sometimes we act deliberately, on the basis of pieces of evidence that we regard as reasons for given actions, but without seeing ourselves as reaching our choices of action mechanically. Our self-

conception includes that strong sense of free and rational agency. To adapt Immanuel Kant's "I think" (*Critique of Pure Reason*, B131), there is an "I do" that must be able to accompany those of my actions that are chosen following systematic reflection, and that should be seen as more than, "This is what my well-stocked and sophisticated brain computed was the most effective way to achieve my goals (including my goal of acting in accordance with my ethical standards)". If we see the "I do" as more than that, without seeing our choices as random or as lacking identifiable sources, then we see ourselves as agents who are free, as well as rational.

As with adoptions of belief, two pointers may help to persuade those who do not think that it is important to see ourselves as choosing some of our actions through deliberation with mastery. The first pointer is the same as the second one that was offered in connection with belief. We do not wish to see ourselves as mere links in causal chains. The second pointer is that if an agent sees himself as having decided on an action through a process that involved a free but controlled choice that was not attributable to antecedents, then he must see himself as owning the action. He cannot see it merely as something that happened with his involvement. I submit that it is satisfying to be required to see our actions as owned by ourselves.

The argument from society

Our social relationships are not relationships between ourselves and objects in the world. We relate to others as fellow subjects. In Peter Strawson's words, we take up a participant attitude toward other people, not an objective attitude ("Freedom and Resentment", page 9). A similar point is made by Carol Rovane, when she elaborates on the essential social role of the recognition of one another as persons (*The Bounds of Agency*, pages 48-49), although she does so in the context of an ambitious project to give an ethical criterion of personhood, defining persons as agents who can engage in agency-regarding relations. In a more philosophically adventurous spirit, it could be argued that a participant attitude would be necessary for a collective consciousness to exist. The argument could, for example, be that a participant attitude would be required in order for the individual to take up the first-person plural perspective that Kay Mathiesen has argued to be the route to the creation of such a consciousness

("Collective Consciousness"). I do not, however, claim that there is in fact any collective consciousness.

The argument from the nature of our social relationships to the desirability of at least implicitly seeing people as deliberating with mastery comes out if we focus on debate, which is the social form of deliberation. Debate may issue in a decision on what to do collectively, just as individual deliberation may issue in a decision as to how to act. Alternatively, debate may be used to try to answer a theoretical question, just as an individual can deliberate in order to arrive at a belief. There may be no need for all of the contributors to a debate to accept the same conclusion, when the aim is to answer a theoretical question, but the desire of contributors to arrive at the truth means that a single agreed answer would still be the ideal outcome.

When we debate with others, we ideally see the others as making contributions that we do not see as accounted for merely by antecedents, whether rational or natural. We see contributions as expressions of the contributors' views, and not merely as expressions of their natures. The difference is one of perception. The views that someone adopts and expresses will in fact be determined by his nature and by the evidence that he has encountered, but to see his views as his considered opinions, rather than as things to which he happens to have been led by virtue of his biological nature and his nurture, is to see them in a new light. That is the light in which we should see them in order to answer to our notion of rational debate. We do not then see ourselves as debating with entities that arrive at their contributions merely by following programs, but as debating with people who exercise mastery over the selection stages of the processes that generate their views. Moreover, seeing other contributors in that light allows us to take up the participant attitude toward them that is the ideal in all of our social relationships. We see them as engaging in debate in the same way that we see ourselves as engaging. We do not see ourselves as mere objects in the world, mechanically reaching and stating our views. And we should not see other contributors like that either.

If we see other contributors as arriving at their views through deliberation with mastery, that has an important effect. If we see mastery, then we do not see views as following mechanically from the interaction of

prior inclinations and the evidence. Consequently, we should not feel entitled to ignore the views of other contributors on the ground that their views merely reflect the influence of prior inclinations on the selection stage. It is useful in a debate to bring evidence to bear, to discuss the weights that should be assigned to pieces of evidence and to dissect the methods that others have in fact used to work from evidence to conclusions, but it is pointless to tell a contributor that he only says what he does because of his prejudices, his past experiences or his financial or other interest in the outcome. We need to recognize the status of contributors as persons whom we see as deliberating with mastery, in order to depersonalize debates while still fully recognizing individual contributors' ownership of, and responsibility for, their views. The supposition of mastery blocks our perception of the inevitability of the influence of how the contributors happen to be, leaving us with their views alone. At the same time, it leaves the views connected to the contributors and to their natures. That is essential to our comprehension of a debate, whether as contributors or as observers. (The one class of debates where this connection would not matter, so that it would not matter if we broke the connection between content and contributors, is the class of debates in mathematics and in the natural sciences where our interest is in the content of the debates, rather than in the history of the disciplines, and where that content is already unambiguous. In such cases there would be no human element, and no hermeneutic work to do. Outside that special class, the interlocking of perspectives that Habermas identifies, and that I discussed in section 1.1, should be taken seriously, even though it may be looser than Habermas allows.)

More generally, if we see others as deliberators with mastery, that helps us to see ourselves as members of a society, rather than as individuals who happen to be surrounded by other members of the species homo sapiens. The latter vision of ourselves would be most unsatisfactory, even if the others did us no harm, or if the whole group was well-structured in a hierarchy like a pack of wolves. It is hard to persuade someone who does not already accept this proposition to agree with it. I can, however, offer a pointer. Would you consider yourself to be a member of a society if its other members regarded you merely as an entity in the causal network,

even one who took account of evidence that appeared to recommend given beliefs and actions but in a way that did not involve mastery over how the evidence was used? Others would then feel free to disregard your views, because they could be attributed to your prejudices or other characteristics, rather than feeling that your views should be considered on their own merits. Going to the other extreme, would you consider yourself to be a member of a society if all of its other members put you on a pedestal and thought that your views should be considered on their merits, but did not expect you to have the same attitude toward their views? Reciprocal recognition that people's views should be considered on their merits has a significant role to play in a society in the modern sense. Groups of people that to varying degrees did not incorporate this reciprocal recognition have existed in the past, and such groups still exist today. They can to that extent be regarded as packs rather than societies, or as societies from which some people, those whose views are not automatically considered on their merits, are excluded by systematic, if sometimes unconscious, oppression. Such oppression is literally anti-social. To the extent that views expressed amount to factual assertions, we can also regard it as a form of testimonial injustice (Fricker, *Epistemic Injustice*, section 1.3).

The foregoing considerations need not deter us from arguing forcefully for our own views, whenever we engage in debate. But if we regard other people's views as worthy of consideration without reference to their origins, while still recognizing people's ownership of their views, that encourages us to limit ourselves to the use of rational persuasion and to forswear manipulation. Rational persuasion means argument that proceeds on the basis that the other party will fully understand what is being said, and in which tricks or disreputable methods, such as straw men or the selective use of evidence, are not used unless by accident, and are admitted if they are noticed. This view can be compared with Carol Rovane's view. She identifies agency-regarding relations, that is, relations in which agents attempt to influence one another but aim not to hinder one another's agency (*The Bounds of Agency*, page 72). But despite the substantial overlap, what she would regard as agency-regarding persuasion differs at the margins from what I regard as rational persuasion. I include in my category attempts to persuade another person to change his rational

point of view, something that Rovane would exclude from her category (ibid., pages 88 and 129-131). I would, however, exclude attempts to undermine someone's integrity to the point where he was left with no rational point of view. Rovane would include in her category hypnosis and deception to which the subject had submitted voluntarily as part of a course of therapy (ibid., page 96), while I would exclude them.

A participant attitude toward those with whom we debate, and a restriction to rational persuasion, do not merely keep debate affable. They can also have considerable political significance, even though politicians themselves are the first to transgress and to use propaganda. If we take up a participant attitude and stick to rational persuasion, that will prevent us from imposing our own ideas of rationality on other people. We will not argue that the modification of their habits of thought or of their specific views through propaganda would be justified because it would be for their own benefit. The limits on the proper conduct of debate therefore have a role to play in blocking the slide from Kantian autonomy to tyranny that was noted by Isaiah Berlin ("Two Concepts of Liberty", sections 3 to 5).

It would be inappropriate to take trouble over debating partners when they were plainly wrong. An expert in a technical field would be quite right to refuse to spend much time debating with someone who stuck to an untenable position, and who showed no desire to be educated. But even then, the expert should choose between attempts at rational persuasion and simply ignoring the opponent. The opponent's views would still be considered on their merits. It is just that it would be obvious to the expert that those views had no merit. There are, however, many topics of debate for which there is no division of people into those who already know the answers, and others who need only ask them to supply the answers. There may be no single correct answers to be found, or there may be single correct answers that are unknown and that are hard to discover. Beliefs as to the answers to such questions are the beliefs that were singled out as half of my topic in section 1.1. The humanities, the social sciences and practical ethics and politics are rich sources of questions to which there are unlikely to be single correct answers. Instead, we can expect each question to have a range of plausible answers, and a wider range of implausible answers. (There are, of course, degrees of expertise on many such questions, and we may be well advised to debate with some people,

while ignoring others who stubbornly advocate implausible answers.) Practically any field of knowledge can yield examples of questions to which there should be single correct answers, but that we are not able to answer and may never be able to answer. There are plenty of questions in mathematics and in the natural sciences that we can formulate precisely but that we cannot yet answer, although we reasonably hope to be able to do so eventually. There are also questions of historical detail that we know have single correct answers, but that we may well never be able to answer, while at the same time we feel the need to reach provisional conclusions because the unknown truth of the matter would colour our understanding of significant people or events.

When we debate a question that has no single answer or no currently accessible answer, a view of people as deliberating with mastery has a special role to play. Not only does it ground the attitude that we should assess other people's views on their merits. It does so more solidly than could be done by our recognizing that we did not ourselves possess a generally accepted answer to the question, so that we should not dismiss other people's answers. We might be wholly convinced of the correctness of our own answer, and might regard its general acceptance as inevitable in time. If we did not see the purveyor of a rival answer as having deliberated with mastery, we could dismiss her answer as arising out of a faulty process at the selection stage, or as being the mechanical result of inappropriate prior inclinations. Furthermore, it is important that we should be able to accommodate the firm holding of our own view, alongside a willingness to consider other views on their merits. When there is no consensus, it is not always appropriate to be tentative and to ground our attitude to other people's views on the thought that they have as good a prospect of being right as we have. It is part of being human that we sometimes have the courage of our convictions. We should ground our attitude in a way that allows for our own conviction, and should not found it on obligatory pusillanimity. At the same time, we should not be so dogmatic as to ignore Cromwell's admonition to "think it possible you may be mistaken" (*The Letters and Speeches of Oliver Cromwell*, volume 2, letter 136, 3 August 1650).

The argument from phenomenology

When someone is deliberating, he may consciously take the view that his

brain merely follows a course that reflects the rational antecedents. In retrospect, he may see how rational antecedents made it most unlikely that he would have reached any other conclusion than the one that he actually reached. If he deliberates out loud, others who know him may note the course of his deliberation and may not be surprised. And an observer who could monitor all of the atoms in the deliberator's brain and the incoming stimuli could predict the course that his brain would take. The deliberator can accept all of that, but when he is deliberating, it does not feel like that at all. He has a clear feeling that the choices of sets of weights to attach to pieces of evidence and of methods of argument that are in general open to him, are all on this occasion open to him, and that any selections that he might make would be made under his control, not at random. That is, it will seem to him, at the time, that he deliberates with mastery, even though others at the time, and he in retrospect, might see the process of deliberation as amounting to the following of a program. This is the argument from phenomenology. If we see ourselves as deliberating with mastery, we have a straightforward way to accommodate the phenomenology of deliberation. I shall set out an additional argument from phenomenology in section 3.2.

1.5 Mastery and other views

In this section, I outline two groups of views that give useful points of comparison. The first group comprises views that, in one way or another, directly incorporate the awareness of free choice. The second group comprises analyses of deliberative processes that allow for considerable sophistication, but that do not take us beyond seeing ourselves as following programs. That limitation will lead up to the challenge that will be faced in chapter 2, the challenge of how to see ourselves as deliberating with mastery.

The awareness of free choice

Mastery over the weighting of pieces of evidence and over their use in deliberation may be compared with the first of the three gaps in the process of deliberation, decision and action that John Searle identifies (*Rationality*

in Action, pages 12-17 and chapter 3). This is the gap between having all of the reasons for choosing each of the options, and making a decision. The agent does not feel that he is driven by the reasons to make a particular choice. They do not appear to be sufficient for the outcome. The other two gaps, one between decision and action during which the agent could change his mind, and the other between the stages of an extended action during which he could decide to discontinue the action, can be seen as variants of the same gap. The agent remains in control. He is not driven onward by the reasons for the course of action that he actually chose or by the fact that he has already made his choice.

This does not mean that Searle and I are saying the same thing, although several of the differences amount to differences of emphasis, rather than disagreement. Searle links the gap directly to freedom of choice of outcome, whereas I think in terms of the subject's mastery over the considerations that bear on the decision, both the available evidence and the subject's prior inclinations. This is largely a difference of presentation, but it is not entirely so. I would not make the strong connections between rationality, free choice and alternative possibilities that Searle makes (ibid., pages 142 and 201-202), unless it was on the clear understanding that this was merely a claim that we needed to see things as if the implications of those connections held good. Searle says that we need to treat the gap as real in order to live as we do (ibid., pages 70-73). While I agree, I concentrate on three specific features of our lives: our self-conception, our social interactions and the phenomenology of deliberation. Finally, my merely claiming that we can and should see instances of deliberation as involving mastery, rather than claiming that mastery corresponds to real features of the world, has the consequence that I do not feel the pressure that Searle feels to align an account of deliberation with the facts of neurology (ibid., chapter 9).

Another view is that of Robert Nozick (*Philosophical Explanations*, chapter 4, section 1). He sets out how we assign weights to the various considerations that might bear on a decision. The assignment of weights on one occasion will influence assignments on future occasions by setting a precedent, but not an inviolable one. Nozick also argues that the assignment of weights on a given occasion can set a precedent under which

that assignment is itself subsumed (ibid., pages 300-301). He addresses the risk of seeing our choices as random and uncontrolled by arguing that a choice can be explicable without its falling under a covering law, and he uses our own experience of choosing in making that argument (ibid., pages 301-306). Nozick's views criss-cross with mine, but our basic approaches are different. He seeks to develop the naturalistic framework, and to use our own experience of choosing in order to give us insights into that developed framework. I shall argue in chapter 2 that we should go beyond the naturalistic conception of ourselves.

Finally, Lucy O'Brien puts forward a particular significance of our standing back and considering the options that are open to us, when she argues that an agent's awareness of her actions results from her "acting on the basis of an assessment of possibilities for acting" (*Self-Knowing Agents*, pages 114-122 and 182-190). The range of actions that is involved is much wider than the range with which I am concerned, actions that are chosen after the agent has stood back and considered both the significance of the evidence and the use that should be made of it. Indeed, O'Brien notes the risk of an over-intellectualist account of an agent's grasp of her basic actions, actions that are performed directly, without doing anything else (ibid., page 186). Furthermore, O'Brien has a different goal from mine. Her goal is to put self-knowledge on a firm footing. There is no reason to think that she would for her purposes need to impute to agents the strong sense of mastery for which I argue.

Sophisticated programs

A leading analysis of deliberation before action is given in chapter 5 of Berent Enç's book *How We Act*. A study of Enç's proposal will show why such analyses are not in themselves enough to allow us to see mastery, even though they may well achieve the aims of their authors.

Enç offers a causal model of deliberation. It relies on event causation, and not on agent causation, so its naturalistic credentials are impeccable. He starts by considering the behaviour of animals in response to threats. Such behaviour can be represented by us in the form of conditionals. If certain sensory inputs are present, take some given evasive action. There is event causation within the animal, but no deliberation. Furthermore, the

operation of such a sequence of causes and effects has nothing much to do with the conditional. The conditional is our gloss on something that is not intrinsically a piece of logical reasoning leading up to a decision to perform the action that is identified by the consequent. The consequent does not get detached as an instruction, or as a recommendation that can then enter into further reasoning within the animal. Enç then argues that we move up to something qualitatively different from such behaviour when the causation and the conditionals come together, with representational states causing behaviour by virtue of their conditional content. Consequences of possible actions can be ranked in order of preference by reference to higher-order priorities, allowing us a subtlety of reasons-based decision-making that takes us far away from automatic reactions, but only if those consequences are themselves represented articulately. Finally, Enç's approach allows him to address the phenomenological feel of free choice. He argues that the role of conditionals in his model requires the agent to have the feeling that she could do any one of a range of different things, a feeling that she could have by imagining herself as adopting each of the options.

Enç addresses the criticism that a naturalistic causal account of action appears to leave no place for the agent (ibid., pages 133-135). He then goes on to tackle the criticism that such an account seems to give an inadequate role to reasons. This second criticism is close to the one that I would make, which is that the available role for reasons is inappropriate, rather than inadequate, because it leaves no room for mastery. Reasons are not seen as operating in the right way for my purposes, purposes that differ from Enç's. A review of Enç's response to the criticism that he considers will show that my criticism could not be addressed by Enç's proposal. This follows from the fact that Enç's description of the process of deliberation leaves us seeing the subject as effectively following a program, despite the complexity of the process that he describes.

Enç accepts that rational action must result from the weighing of pros and cons of possible actions (ibid., page 136), and I have no quarrel with that, although others do. Markus Schlosser disputes that acting for reasons necessarily involves deliberation (*The Metaphysics of Agency*, pages 189-193). While Schlosser may be right, there are some types of action, such as

the action of agreeing to a business plan, which must, if undertaken for reasons at all, involve the weighing of pros and cons. If I can show that seeing mastery would require more than proposals such as Enç's could offer us in relation to those types of action, that will do for my purposes.

The kernel of Enç's response is his claim that we can account for deliberation on pros and cons in naturalistic terms, provided that some of the states of the deliberator are representational, and that those representational states play their causal role by virtue of their conditional content, content on the lines of "If I do this, then that will happen" (*How We Act*, pages 136-137). The agent's conscious responsiveness to conditional content is certainly significant. But Enç's approach merely has the subject work through the consequences of alternatives and select the most desirable option. Once all of the facts about consequences and preferences are in, the computation proceeds as if merely following a program. My concern is that this does not allow us to see deliberation with mastery.

Enç recognizes the need for variability in weightings (ibid., pages 222-223), but he does not go into detail as to the mechanism of variation. Indeed, he appears to rest content with the variation of weightings in response to circumstances. Fortunately, a possible mechanism is available. Harry Frankfurt distinguishes between first-order desires, such as the desire to take a dose of a drug and the desire to be free of addiction to that drug, and second-order desires, such as a desire that the desire to be free of the addiction should prevail ("Freedom of the Will and the Concept of a Person", pages 327-329). Such a hierarchy of desires certainly allows for flexibility in weights and in the use that is made of pieces of evidence in argument. It also allows for the production of reasoned arguments as to why certain choices were appropriate. In short, it gives us most of what we want, and it can do so within a naturalistic framework. It is true that Enç's diagrams of decision-making processes include boxes labelled "higher order of preferences", but these appear to be general preferences that allow the evaluation of consequences, rather than second-order desires in Frankfurt's sense (Enç, *How We Act*, pages 157, 158 and 171). It is not clear whether they could lead to the flexibility to which second-order desires could lead. But it does not matter for my purposes whether either Frankfurt or Enç has the correct approach, because what they offer is still not enough.

It is not enough because the addition of a layer of higher-order desires, or even several layers, would merely make the program more complex. Seeing a subject as deliberating with mastery involves seeing him as standing above any program. Making a program more elaborate cannot give us what we need. For the same reason, the two-level theory of the mind that is given by Keith Frankish, in *Mind and Supermind*, cannot give us what we need. (Frankish distinguishes a special category of supermental states and processes, which are of general application and of which we can be consciously aware. We can use these states to regulate lower-level mental processes of which we are not generally conscious, and to resolve tangles that afflict those lower-level processes.) Not even the rich and sophisticated model of self-governance that Michael Bratman develops in *Structures of Agency* will suffice.

In order to see ourselves as deliberators with mastery, we do need the sort of complexity of structure that is offered by proposals such as those of Enç, Frankfurt, Frankish and Bratman. We also need the meta-symbolic ability, the ability to symbolize our symbols, that is described by Hans Lenk in "Humans as Meta-symbolic and Super-Interpreting Beings". Without that ability, we would not be able consciously to evaluate our lower-order preferences by reference to our higher-order preferences. These things are necessary, but insufficient. With the complexity that they allow, we can see ourselves as attaching weights to pieces of evidence, as putting the weighted pieces of evidence into play and as selecting methods of argument that take us from weighted evidence to conclusions as to what to do. We can also see our selections as made under the guidance of higher-order principles, rather than their being random or without identifiable sources. We can then see that the choices of weights and of methods of argument can be our choices. They are choices that make sense and that can be justified. We can even see our choices of higher-order principles in the same light, by taking those choices to be made under the guidance of yet higher-order principles. We are well into the realm of which Kant wrote, when he stated that "only a rational being has the capacity to act according to the representation of laws, that is, according to principles" (*Groundwork of the Metaphysics of Morals*, Ak.4: 412). We also have enough to give us the reflective distance between our incentives to act and our decisions that Christine Korsgaard

identifies (*Self-Constitution*, section 6.1.7). At least, we can have that reflective distance so long as "incentives" has its ordinary meaning, and is not read so broadly as to encompass all of the rational antecedents of our deliberations.

Building in all of this brings us almost to the point where we can see ourselves as deliberators with mastery, so that we can see ourselves neither as followers of programs, nor as lacking control. But it does not quite get us there, because when a subject gets above the top of the hierarchy of principles, he stands above the program but is no longer guided entirely by the highest-order principles, so that the process is seen as being at least to some extent out of control. On the other hand, when the subject is guided entirely by the highest-order principles, he no longer stands above the program. A position of mastery that was thus conceived would be unstable. A subject might see himself as above the highest-order principles, but he could at best only see himself as sustaining that position for an instant. He would need those principles to catch up with him, in order to maintain control. (This would not imply an indefinitely large number of layers of principles, with a new layer being added at each instant. The principles could catch up by the subject's sinking back down.) Even this momentary mastery might be impossible to conceive coherently.

The instability at best, and incoherence at worst, of a position of mastery as thus conceived is highlighted by an argument that is put forward by Simon Blackburn (*Ruling Passions*, chapter 8, section 3). He argues that when we deliberate about what to do, we look outward at the world rather than looking inward at our desires and inclinations. We should not see ourselves as pushed around by our desires and inclinations, but should welcome their influence because they constitute us, or at least the more stable ones do. They are not external forces. Likewise, Christine Korsgaard's vision of a subject who adjudicates between desires, rather than being carried along by the strongest desire, is a coherent and stable vision because the subject has contentful principles according to which he makes his decisions, albeit principles that he has in his view legislated for himself (*The Sources of Normativity*, page 100). Correspondingly, it would seem to be a mistake for a subject to see an internal queen who would sit above all of his desires, inclinations and principles and weigh them in the

balance, as well as weighing the relevant evidence about the state of the world, when working out what to do. That would seem to be, in Leibniz's words, "une prosopopée ou fiction un peu mal entendue", "a prosopopoeia or fiction a little ill-conceived" ("Remarques sur le Livre de l'Origine du Mal, publié depuis peu en Angleterre", section 16, in *Essais de Théodicée*). We may add a standard objection to anything like the picture of a queen who sits above all rational antecedents, which is that the queen would lack any character, so that her decisions would be random. I shall return to this objection in section 3.8. My response will be to keep the queen's operation mysterious, so that her decisions cannot be characterized either as random or as non-random.

If we seek an account of what we usually do, Blackburn is right. We take our desires and inclinations as given. Sometimes we review our first-order desires in the light of our higher-order desires, along the lines that have been proposed by Frankfurt, but Blackburn need not find that objectionable. What he does find objectionable is the idea of weighing all of our desires and inclinations in the balance, along with the evidence that we have. He goes beyond Leibniz's point, by arguing that if we are concerned with what to do in a given situation, the last thing we want to do is divert our attention from the immediate problem to our thoughts about it (*Ruling Passions*, page 254). Blackburn's point could be disputed whenever the problem was of a type that made systematic deliberation appropriate, and whenever we wanted to be wary of the influence of our prejudices, but in any case the point does not bear on my model of deliberation with mastery. I am only concerned that we should see the subject as standing back and freely choosing how to weight the evidence and which method of argument to use. The subject can be seen as not merely following a program that captures the content of the rational antecedents, without being seen as thinking about his prior inclinations. Even so, the other difficulties stand. We must look for another way to see mastery.

CHAPTER 2

How to see mastery

If we should see ourselves as having mastery, we must understand how to see ourselves in that light, and the implications of doing so. This chapter covers that ground.

In section 2.1, I baldly assert that we should simply see mastery, and note that in order to make this assertion acceptable, a fuller picture must be developed and must be related to our scientific view of the world. In section 2.2, I set out the relationship between deliberation with mastery and the causal closure of the physical. The latter must be overlooked in order to allow us to see the former. This raises the question of how we are to see deliberation as being appropriately controlled. In section 2.3, I introduce a way to supply the required control. In section 2.4, I set out two conceptions of humanity, one of which is the right one for my proposal. Setting out these conceptions highlights the non-naturalistic nature of my proposal.

2.1 The assertion of mastery

My proposal is simply to see a subject, when he deliberates, as having mastery, in that he has control and does not merely follow a course that is indicated by the rational antecedents. We need not see the subject as bound to follow such a course, even if we ignore the option of flouting the accepted logic and also ignore the possibility of a failure to implement that logic correctly in grey matter. We can see the subject as making free, but controlled, choices at the selection stage of deliberation. Those choices have an identifiable source in the subject.

Suppose, for example, that someone is always inclined to discount consequences of an action that lie more than five years ahead, because of the uncertainty of such distant consequences. Suppose in addition that in fact he always adheres to that policy, but that he is, so far as his general capacities go, able to consider breaking with the policy. Then on the approach that is being proposed here, he should be seen as having the option, on each occasion of deliberation, of breaking with his policy, becoming less concerned about uncertainty and starting to consider consequences that lie in the distant future, without flouting the logic of deliberation that is taken as given. Furthermore, we should not suppose that any lack of control would be involved in any such deviation from the normal policy that might take place. (It may not actually take place. We are only considering how we should react to a deviation that we are to see as possible, even though physical determinism may mean that it is not in fact possible.) Nor should we see control as being supplied by any higher-order policy, because we need to see the subject as having freedom to go outside the program that would capture the entire content of the rational antecedents. Thus I propose a flat denial that the subject's progress toward a conclusion should be seen either as the working out of a mechanical process, or as random and uncontrolled. Even when there is no change of policy, no going outside the program, we should see the subject as freely deciding to proceed in a way that happens to amount to compliance with the program. A subject's change of policy, suddenly considering consequences ten years hence when he had previously only ever considered consequences that lay up to five years ahead, would not in everyday life be likely to disturb us. We accept that people sometimes change their minds. The current discussion is simply meant to expose what is involved in our regarding such a change of mind as always an option for the subject, without regarding it as an option, the exercise of which would amount to the operation of any kind of mechanism that governed the thought of the subject and that would ensure that he took up the option, or the exercise of which would have no identifiable source.

My bald assertion that mastery should be seen will naturally give rise to a suspicion that my proposal has the advantages of theft over honest toil. To defend the proposal against that charge, I must develop a full

picture of ourselves as exercising mastery, and must establish that this picture gives us useful results, while also being satisfactorily related to the picture of ourselves and of the world that the natural sciences give us. These tasks will occupy me in this chapter and the next.

2.2 Causation and control

The place to start is the relationship between seeing people as deliberating with mastery, and seeing them as physical beings who are embedded in the causal network. We are to see the subject as having a range of options. She is to be seen as able to choose any one of a range of courses of deliberation, including courses that are not indicated by the rational antecedents, so that she could go outside the program that captured their content. But we cannot see the subject as having such a choice, even if we do not see her as exercising it, if we take full account of the determinism that we see at the level of physics. A link between determinism and compliance with the program is given by the fact that our processes of reasoning and the courses of development of our brains do generally march in step. Given that we generally reason in accordance with our rational antecedents, the deterministic course of development of a subject's brain will in general correspond to compliance with her program. This link only holds most of the time, not always, although it could be argued to be necessary that it does hold most of the time. But the fallibility of the link will not matter here, because the argument will be based on the fact that if mastery is seen on a given occasion, a wide variety of options must be seen as open to the subject on that occasion. It will only matter that determinism in the course of development of the subject's brain would not allow such a range. It will not matter whether or not the course of development that was physically possible would correspond to compliance with the program.

References to determinism might appear to assume too much. Our physical theories also allow for randomness. But that randomness is irrelevant in this context, for two reasons. First, if the randomness is based on quantum mechanics, even a single brain cell is too large and too warm for quantum effects to be of any significance. Second, we cannot expect

randomness to turn up obligingly at appropriate moments, to give us the choices that we want to see ourselves as having. This second reason for not relying on appeals to randomness applies not just to randomness that is based on quantum mechanics, but also to randomness that is taken by neuroscientists to arise at higher levels. Examples are when synaptic transmitter release is modelled stochastically and, to move up to an even higher level, when the successive states of activity of a network are modelled as a Markov chain (Dayan and Abbott, *Theoretical Neuroscience*, pages 178-180 and 273-276). Likewise, we cannot rely on the lack of strict psychophysical laws, laws that might rigidly link mental and physical descriptions of the same events, a lack for which Donald Davidson argued ("Mental Events", part 2). There might be psychophysical laws that were not loose enough, or that were not loose in the right ways, even though they were not perfectly strict. I shall also make some points about the lack of help from the multiple physical realizability of mental states below. In short, we must face determinism squarely. But the claim that this is a problem still requires an argument in its support.

The precise claim is that we must, when we see the subject as having mastery over the selection stage, overlook the causal closure of the physical. We must do so even though the outcome of the selection stage will never in fact differ from any outcome that would be predicted by an analysis of physical processes, nor should we ever expect it to differ from such an outcome. We are to overlook what the natural sciences tell us about causal processes, not fly in the face of the predictions that the sciences allow us to make.

The doctrine of the causal closure of the physical must first be defined. I shall take it to be the doctrine that "every physical phenomenon that has a sufficient cause has a sufficient physical cause" (Montero, "Varieties of Causal Closure", page 174). This formulation does allow that some physical phenomena might lack sufficient causes, and that possibility needs to be excluded in order for my argument to go through. We can exclude it because we are not in the realms in which quantum effects are relevant, and because there are no other specific reasons to think that physical events might just happen. It is true that even though an event of the type with which we are immediately concerned would not be influenced

by quantum effects, it might have had ancestors in its causal history that were so influenced. But while that would make space for a lack of sufficient causes for the event with which we were concerned, there would be no reason to think that there would have been a gap in physical causal history that could have accommodated a mental cause, a possibility that would create problems for my argument. The gap would have been too small for anything that had determinate mental content to fit into it. It could not have accommodated enough information. And there would be no reason to think that any appropriate non-physical cause would have been around at the time when the gap existed, ready to take advantage of it. Finally, if we combine the exclusion of a lack of sufficient causes with the doctrine of the causal closure of the physical, that excludes causal action of the non-physical on the physical, or at least makes it idle. This exclusion is also needed in order for the argument to go through.

Three facts together mean that we must overlook the causal closure of the physical in order to see the subject as having the options that we want to see her as having. The first fact is that the brain is embedded in the causal network, just like any other object. The second fact is that there cannot be a difference in mental states unless there is a difference in physical states. (The difference in physical states might be inside or outside the subject. I do not wish to dismiss externalism about mental content.) Any scientifically plausible account of the human mind must make it dependent on the physical states of the world, and primarily of the brain, to that extent. The third fact is that the correspondence between those physical states and the mental states of a given person must be stable in the physical-to-mental direction. If it could vary from moment to moment, we would have no explanation of how we were able to engage in extended processes of reasoning. That ability needs to be explained by the existence of a reasonably stable, although not necessarily immutable, function from the states of a person's brain and of the surrounding world to her thoughts. Only a physical brain has the inertia that can keep a train of thought on track, and the effects of that inertia must be transmitted to the train of thought via a stable correspondence. It follows that a given course of development of a subject's brain is only going to correspond to one course of reasoning, and that the correspondence must be settled in advance. It

cannot be varied to any significant extent as the reasoning progresses. I shall now explain why these three facts require us to overlook the causal closure of the physical.

We are to see the subject as having open to her a range of possible sets of weights to attach to pieces of evidence, and a range of possible methods of argument. The making of one selection, of weights or of method, rather than another, would require the course of development of the subject's brain to remain appropriately aligned with the subsequent course of reasoning, even if the differences between the possible selections, and hence the differences between the possible subsequent courses of reasoning, were considerable. It would only be possible for the course of development of the subject's brain to remain appropriately aligned with the course of reasoning, after the making of one or another of a range of significantly different selections, if one of two alternatives applied. The first alternative would be for the choice between the selections to be governed by some physical process that would also steer the brain into continued alignment with the course of reasoning, whatever choice was made. The second alternative would be for the choice to steer the brain. The course of development of the brain would have to take a sudden turn. This would require a deflection of the physical world from its course by something that was purely non-physical, in the sense that it did not even have a physical counterpart with which it was correlated for reasons other than action of the non-physical on the physical. If there were such a counterpart, we would have an instance of the first alternative. As I am about to argue that the first alternative would be unacceptable, rather than useless, such physical counterparts of the non-physical items under the second alternative must be avoided.

The first alternative would be unacceptable because physical determinism, the causal closure of the physical, the physical nature of the brain, the dependence of the mental on the physical and the stability of the correspondence between physical and mental states in the physical-to-mental direction would mean that we could not see the subject as having open choices. The physical process could only go in one way, so the choice between selections could only go in one way. We must therefore adopt the second alternative and see the choice as steering the brain, that is, see the

purely non-physical as acting on the physical. But we cannot see that as happening, so long as we have the causal closure of the physical in view, and also require physical phenomena to have sufficient causes. I conclude that if we are to see deliberation with mastery, we must overlook the causal closure of the physical. But we must not expect ever to see an actual deflection of the physical world from its causal course. We will never need to find actual action of the non-physical on the physical, in order to explain the alignment between the course of development of a subject's brain and the course of her reasoning.

Where does that leave the indeterminacy of reasoning that we can often expect to be allowed by programs, as distinct from physical indeterminacy? The playing out of indeterminacy of reasoning would not satisfy our desire to see ourselves as making free and controlled choices. And as it happens, its playing out would not require us to see anything that conflicted with the causal closure of the physical. Indeterminacy at the level of reasoning need not be underpinned by anything strange at the physical level, not even physical indeterminacy. Indeterminacy in processes of reasoning could, for example, arise out of the subject's use of loose definitions of concepts, while the use of a specific set of definitions of concepts, whether tight or loose, could correspond to any one of a wide range of different brain states. Each initial brain state could participate in an entirely deterministic evolution of the physical world, but each different initial brain state would lead to a different path of physical evolution, those paths corresponding to potentially (but not necessarily) different outcomes of the process at the level of reasoning. The non-existence of psychophysical laws that are strict in the mental-to-physical direction, a non-existence that allows each mental state to be correlated with any one of a range of possible physical states, is a necessary condition for indeterminacy at the mental level on the back of determinism at the physical level. It is necessary because if there is no difference between two states of the world at the physical level, there can be no difference at the mental level. And the causal processes in question are deterministic. Quantum effects are irrelevant in this context because of the size and warmth of brain cells. Nor are causal processes indeterministic by virtue of the need to model certain brain processes stochastically. That need

merely reflects our inability to handle detailed information that captures the full physical complexity of a large system.

There is another necessary condition, which is that the indeterminacy should be built in at the level of reasoning. It may, for example, be built in through the subject's use of loose definitions. The causal mechanism cannot on its own, without looseness that is given at the level of reasoning, generate the same effects as would result at the level of reasoning from, for example, using loose definitions. The causal mechanism obviously could not do so in the absence of multiple realizability. Adding multiple realizability does not do the trick, for two reasons. The first reason is that the physically possible courses of evolution of the brain from its possible initial states would be most unlikely to correspond to the logically possible courses of loose reasoning, unless that range of initial states of the brain was primarily defined by sources of looseness that were given at the level of reasoning, such as loose definitions. The second reason is that multiple realizability need not in itself open up a range of possible paths at the level of reasoning. We can see this by considering a process of reasoning that uses tight definitions and a precise logic, eliminating indeterminacy at that level. Such a process can still be realized in a variety of different ways at the level of brain cells. Feedback loops and other control mechanisms will, however, ensure that all of the different evolutionary paths, paths that start with the different initial states of the brain, include only states that correspond to stages in the correct path at the level of reasoning. At least, they will ensure that happy outcome absent malfunctions of the brain. The fact that our brains can keep themselves on track like this is one aspect of their ability to handle the processes of reasoning that we use.

These arguments about loose reasoning have parallels in arguments that are directly related to mastery. There are two good reasons why we should not expect multiple realizability to allow us to accommodate mastery over the selection stage, without overlooking the causal closure of the physical. The first reason is that it would be a miracle if the range of possible courses of physical evolution of the brain corresponded to the range of choices that must be seen as open to the subject. The second reason is that a brain can follow determinate courses of reasoning, despite being liable to realize any given course of reasoning in any one of a variety

of physically different ways, so that multiple realizability need not open up the options that we might hope it would open up.

Those who do not believe in suspension of the causal closure of the physical might be tempted to reject my entire project at this point. That would be premature. I do not myself believe that such suspensions actually occur, nor do I believe that there is a special form of causation, such as agent causation, alongside ordinary physical causation. The project will be saved from the objection that causal closure is never suspended by the fact that the proposal is only to see deliberation as if the causal closure of the physical were not there. There is no claim that it is in fact not there. But there is still a debate to be had about the legitimacy of overlooking the causal closure of the physical. It would not be legitimate if it required us to contradict the results of the natural sciences. Overlooking causal closure does not require us to contradict empirical data at the level of measurements that can be taken, even if we do find sufficient physical causes for physical phenomena whenever we bother to look for them. Just for once, Hume's point that causal necessity is not straightforwardly visible is helpful (*An Enquiry Concerning Human Understanding*, section 7, part 2). Nonetheless, the principle of the causal closure of the physical is on the border between being physical and being metaphysical. The more that it can be regarded as metaphysical, the less the danger that overlooking causal closure would amount to contradiction of what the natural sciences told us. But even if that defence were to fail, I could derive support from the legitimacy of partial descriptions of the world, from which physical details were omitted. I discuss that legitimacy in section 3.6.

If we have to overlook the causal closure of the physical in order to see a subject as deliberating with mastery, we have a problem of how to see our processes of deliberation as subject to appropriate control. We do not want to see ourselves as mere ciphers, without prior inclinations, whose deliberations do not reflect our own characters. Nor do we want to see our choices of action or adoptions of belief as lacking sufficient identifiable sources. The existence of a logical course that can be described as the execution of a program, and that can be physically embodied, addresses the need for control very well. But if the subject were seen as free to go outside the program, not by flouting the accepted logic but by breaking free from

the effect of her prior inclinations, if the underlying physical causal mechanism were not seen as all-powerful, and if no new source of control were brought into view, could that be satisfactory? My proposal does not require us to ignore the influence of prior inclinations. We can see them as having influence, subject to due consideration. But if that consideration were to be seen as uncontrolled, there would be a danger that we would have to see the process that was influenced by the prior inclinations as uncontrolled too. We might have to say that a selection stage could have turned out differently, leading to its being followed by a different calculation, which would in turn have led to a different substantive outcome, for no particular reason. The consequent picture of ourselves would not measure up to our self-conception.

2.3 Coping without causation

A consequence of the need to overlook the causal closure of the physical, and the need to tackle the problem of control, is that the best approach is to see human beings, when engaged in deliberation, as points of origin. Their decisions at the selection stage on how to weight evidence, and on how to argue, should be seen as originating in themselves at times of decision, with no prior causal or computational history, rather than as consequences of rational or natural antecedents. There is no suggestion that rational antecedents should be disregarded by the subject. Indeed, pieces of evidence may by their nature suggest weights. Thus one piece of evidence might be that information from a given source was generally reliable, and that might lead to significant weight being given to other pieces of evidence that were derived from that source. But the substantive choice of action or the adoption of the substantive belief, the final product of a process that is seen as one of deliberation with mastery, cannot simply be seen as having emerged mechanically from the antecedents, whether rational or natural. We need to see a special contribution from the subject. This is the free selection of the set of weights to attach to pieces of evidence, and of the method of argument. In that sense the subject is to be seen as a point of origin, as the identified source of the free decisions. This

does not mean that the subject's special contribution should be seen as making any practical difference. We must acknowledge that the outcome would have been the same, and that we would have anticipated the same outcome, even if we had not seen this special contribution as having been made. We know that the subject is part of the physical world, and that the world is not in fact going to veer off its causal paths, save on account of mere indeterminacy.

Decisions that are seen as originating in the subject at the time of decision, and not as following from antecedents, will be regarded as the results of a process that I shall call subject origination. Decisions as to weights and methods will be seen in that light when we see the subject as having mastery over the selection stage. When a human being engages in subject origination, I shall speak of her being a point of origin. I shall use the term "originator conception" for the conception of human beings as such points of origin.

So far, the problem has been re-labelled rather than solved. In chapter 3, I shall say more about subject origination. I shall also set out a concept of the subject under which we can see ourselves as engaging in subject origination. But I shall make a start here with a brief description of how the process is envisaged, and of how instances of it should be seen as related to the causal network of the world.

We are to overlook the causal closure of the physical. We are to see a causally efficacious extraneous intervention at the selection stage of deliberation. The supposed causal effect is the steering of the brain into a particular course which will then match the course of reasoning that follows from the subject's free but controlled choices of weights and of method. The causal efficacy of such an intervention would have no place in nature. But we need to see an intervention if we are to see the subject as freely choosing from the full range of sets of weights to attach to pieces of evidence, and from the full range of methods of argument, that would in general be available to her. And the intervention needs to be seen as causally efficacious, in order to explain how the brain then keeps up with the mind. The fact that the intervention would, if it were real, be causally idle, because the world never in fact deviates from its normal causal paths, does not remove the need for it to be seen as efficacious.

This overlooking of causal closure is not a matter of imposing dead ends on causal chains just before the moment of decision, which is what would be needed if the subject were seen as entirely within the causal network. Instead, it is a matter of seeing the subject as able to act from outside the causal network. There is also no suggestion that we should elaborate the picture into one that involved a homunculus within the subject, tempting though that might be. The introduction of such a creature might seem to be an easy way to give the subject control. But it would only generate a regress, when we asked who controlled the homunculus, or land us back where we started, if the homunculus were seen as embedded in the causal network. My proposal does not involve seeing as inserted into the causal flow anything more than the emanations of the subject, whose nature as the source of those emanations is not further specified. The originator conception of human beings is the conception of them as points of origin of such emanations. Whenever we see a human being as deliberating with mastery, we see an instance of subject origination. Whether the consequences are as envisaged by the subject should, of course, be seen as depending on the working out of the many causal currents that started before the deliberation, as well as on the subject's input. The new causal current that is seen as starting with the subject merges into the stream of the world, like a tributary flowing into a river. What happens downstream of that point will be very greatly affected by the flow of the main river upstream of that point.

On the specific point of control, I must admit to the issue of a promissory note that is not backed by any gold, even if my proposal might not quite amount to theft. No mechanism of subject origination is specified. It would not even be possible to specify one while retaining the function of subject origination, because the efficacy of any mechanism could only be explained if it, or some other mechanism, the states of which were for good reason in a stable correspondence with its states, were seen as embedded in a causal network. Such embedding would be needed in order to account for the embedded mechanism's functioning, that is, for the efficacy of its parts as those parts acted on one another. Thus I assert, but do not account for, the subject's control over the selection stage of deliberation. My defence against a charge of theft is that I only propose

seeing people as engaging in subject origination. I do not assert that there is any real process of subject origination.

In the same vein, I should emphasize that seeing deliberations as conducted with mastery, and hence as involving subject origination, does not entail commitment to the proposition that different choices could in fact have been made, that is, to the principle of alternate possibilities that has been proposed as a litmus test both of freedom and of moral responsibility, although its status as a litmus test of moral responsibility has also been attacked (Frankfurt, "Alternate Possibilities and Moral Responsibility"). Such a commitment is avoided because we merely adopt a way of looking at processes of deliberation.

2.4 Two conceptions of humanity

I have introduced the originator conception of human beings. In this section, I shall contrast it with the naturalistic conception. This will bring out the non-naturalistic nature of my proposal.

Applying the naturalistic conception, we see ourselves from the outside, objectively. In practice this means that we see ourselves simply as animals, physical beings in the world who are subject to all of the natural laws of the world, and as vehicles of processes within the life of the mind, including processes of selection and of calculation, that are embodied in our grey cells. We write natural histories of ourselves. We note that it is perfectly possible that everything that we think and do is determined, even though multiple layers of processing within our brains give us a remarkable subtlety of thought and of conduct. If we apply the naturalistic conception, we can perfectly well adopt the program view of our deliberations.

Applying the naturalistic conception means describing ourselves in terms that only bring out features that would be visible to a very wide range of rational beings. Such features include not just physical features, but also features that are given in terms of information theory. In broad terms, they are the features that could be picked out by scientific theories, including the theories of psychology. (I shall discuss the extent to which intentional descriptions could be used in section 3.7.) The point of singling out the

naturalistic conception is that when we apply it, the accounts that we give are uncontentiously intellectually respectable. Their widespread accessibility, well beyond the human race, testifies to that respectability. We can apply the naturalistic conception without having to justify our doing so.

The reference to a very wide range of rational beings involves a vagueness that is sadly unavoidable. It is tempting to eliminate the vagueness by referring to all rational beings, but that would go too far. The conduct of some of the natural sciences, including those that encompass the scientific study of humanity, requires use of the concept of causation in a form that incorporates the notion of causal necessity. As will emerge in section 6.4, we cannot be sure that this concept would be shared by all rational beings, although we can assume that it would be shared by a range of possible or actual rational beings that was much wider than humanity alone. The use of intentional descriptions would further narrow the range of rational beings that could comprehend our naturalistic descriptions of ourselves.

Applying the originator conception, we see ourselves as points of origin of our decisions as to how to weight evidence and as to what methods of argument to use, rather than merely as followers of programs. This reflects our own sense of deliberation. Correspondingly, we want to see more than would be allowed if we saw ourselves merely as the loci of complex interweavings of many causal chains. We would have no reason to see ourselves as any more than such loci if we saw our brains merely as implementing programs that captured the content of the rational antecedents of deliberations. It is, however, interesting to note a tendency among some philosophers who, like me, seek a picture of ourselves that we can accept as true to our lives, to take as a starting point the idea that many causal chains come together in each human being. One example is the discussion of activity, initiation and guidance control in Fischer, "Compatibilism", sections 6 to 9. Another is the point that is made by E. J. Lowe, that the unity of an action is only reflected in the agent, and is lost in the ramified causal chains that stretch into the past, even when we only go far enough back to see those chains as running all over the different parts of the agent's brain (*Personal Agency*, section 5.4). Lowe does not, however, make his point in order to rest content with a straightforwardly

naturalistic account. He makes it as part of an argument for a non-Cartesian substance dualism.

Application of the originator conception to ourselves brings out apparent features that would not be visible to nearly so wide a range of rational beings as the range that could appreciate a natural history of humanity. A rational being without inner experience that was sufficiently like our own would not appreciate either our attitude toward ourselves, or our inner experience of free but controlled deliberation, and could at best only apply our originator conception to us in a purely formal, contentless way. (Such a being might, however, apply something analogous to the originator conception to himself and to others who were like him. And he might also try to apply his conception to us, while suspecting that this would misrepresent our inner experience.) Application of the originator conception takes us beyond the bounds of natural histories of ourselves. Its application also requires us to overlook the causal closure of the physical, a closure that is taken for granted in the natural sciences. For both of these reasons, the originator conception is a non-naturalistic conception. Any claim that it is legitimate to use the originator conception requires argument.

The naturalistic conception of ourselves cannot be discarded, or even played down. The scientific enterprise, which is the primary field of application of the naturalistic conception of everything, ourselves included, has been so successful in helping us to understand and manipulate the world that it must be a correct way of approaching the world. It is probable that much of the detail of current scientific knowledge gives us final and correct answers on the subjects with which the natural sciences deal, even though in some areas, including fundamentals such as the nature of matter, new revolutions may well lie in store. But if we were to limit ourselves to the naturalistic conception, we would be unable to use the notion of deliberation with mastery to support our self-conception and to accommodate the phenomenology of deliberation. We would lose an opportunity to construct a philosophy that would help to make us feel at home in the world.

The Davidsonian tradition

I have now said enough to be able to comment on the relationship between

my approach and a strong philosophical tradition that would oppose the view that acknowledging the causal closure of the physical would debar us from simultaneously seeing ourselves as going through rational processes of choice of action. This is the Davidsonian tradition that reasons can themselves be causes, albeit in the way of singular causes, the operations of which are not apt to be brought under rigorous covering laws that would be of much use in prediction (see Davidson, "Actions, Reasons, and Causes", and the large literature that takes its rise from that paper). A significant paper that is broadly within that tradition, while not wholly agreeing with Davidson, Jennifer Hornsby's "Agency and Causal Explanation", starts by considering two points of view, just as I adopt two conceptions of ourselves. Hornsby argues that an action can be seen from a personal point of view as a person's doing something for a reason, and from an impersonal point of view as a link in a causal chain (ibid., section 1).

The difference between the Davidsonian tradition and my proposal is this. Working within the tradition, we come to each instance of human conduct, apply the two points of view to it, and then seek to reconcile them. I propose choosing one or the other conception of ourselves first, according to our purposes, then going out and looking for instances of deliberation to view under the chosen conception. Davidsonians seek local reconciliations. I seek a modus vivendi between two global conceptions. This difference is easily explained. Davidsonians are primarily interested in why choices are made, and specifically in explaining the relationship between the outcomes of causal processes and the outcomes that make sense in the light of the reasons of which agents are aware. I am primarily interested in the process of deliberation. My concern is with supposed freedom rather than with actual rationality. For a process to be seen as one of deliberation with mastery, the causal closure of the physical must be banished from the scene from the outset. Davidsonians and I need not disagree. We simply have different concerns.

CHAPTER 3

Subject origination

In this chapter, I argue that the concept of subject origination can do the job that it has to do. I also compare my approach with some approaches of other philosophers. In section 3.1, I locate subject origination and the originator conception in relation to the naturalistic conception of ourselves and in relation to agent causation. Section 3.2 covers the phenomenology of subject origination. Section 3.3 covers both the need to be able to enter into one another's heads, and what is involved in doing so. In section 3.4, I consider the range of beings to which we would attribute subject origination. Robots and aliens are both likely to be excluded, but for different reasons.

In section 3.5, I argue that the concept of subject origination is sufficient to do its job. There are two types of sufficiency at stake, sufficiency of content and sufficiency of status. Content is supplied from elsewhere. The status in question is that of being a mere way of looking at our processes of deliberation. I consider two specific aspects of the work that the concept of subject origination might do, the explanation of human actions and the attribution of responsibility for those actions. In section 3.6, I consider the legitimacy of use of the concept of subject origination. The relationship between the view of ourselves that is given when it is used and the scientific view is crucial here. I also consider the standards which ensure that the view that is given is not an arbitrary view.

In section 3.7, I consider other approaches that are in play in the same field as subject origination. Agent causation differs from subject origination primarily because its proponents mostly seek to establish the existence of something that is in the world, alongside event causation. The intentional stance, as proposed by Daniel Dennett, contrasts with the

originator conception because it could be adopted while conceiving of humanity in purely scientific terms. There would, however, be significant difficulty in arriving at the correct intentional stance to adopt without drawing on our own inner experience. In section 3.8, I propose a concept of the subject that allows us to see ourselves as engaging in subject origination.

3.1 Subject origination and other views

If we adopt the naturalistic conception of ourselves, that gives us a view of each process of deliberation. We see the process as a series of events alongside other events in the world, all of which fall within the network of causes and effects. (If we characterize the process at the level of reasoning, we do not see physical causes and effects, but we do see a process that is tied to a progression of physical causes and effects.) If we adopt the originator conception of ourselves, that gives us a different view. There is a supposed causally efficacious extraneous intervention at the selection stage, and the outcome of that stage is seen as explained by more than the operation of a program. When we see the process in that way, we see it as involving subject origination. We then see the subject as a point of origin. The term "subject origination" is a convenient way of bringing out what we need to see, the extraneous intervention. But its use, in addition to use of the term "deliberation with mastery", not only marks this specific way of supporting our self-conception. It also leaves open the possibility that subject origination might be seen in contexts other than that of deliberation.

The view of human deliberation that is afforded by the naturalistic conception is likely to be thought of as the more substantial view, and rightly so. It sets our deliberation in the context of scientific theories that hold over a far wider area than that of human life. Psychology, in the form in which it applies to us, may be a discipline that can only be applied to human beings, but our psychological nature depends very largely on our biology, and human biology has a great deal in common with the biology of other animals. Biology is in turn intimately bound up with chemistry, and chemistry with physics. Our theories of chemistry and of physics, to

the extent that we have got them right, apply across all or most of the Universe, and they apply both to the animate and to the inanimate. Once we see how the scientific view of our mental lives is set within this much wider context, we can see that it is much more substantial than any view that lacks such connections to our view of the rest of the world.

The view of human deliberation that is afforded by the originator conception lacks this degree of integration with the rest of our knowledge. There is no direct link with independent areas of our knowledge, such as physics and chemistry. There are direct links with history and with the other humanities, in which we see people as rational agents who decide what to do and then act, and in which we see art and literature as capturing our inner sense of living and loving. But as the word "humanities" suggests, these disciplines are not independent of our conception of ourselves as beings who have distinctive characteristics, simply by virtue of being human. In the light of this lack of direct links with independent disciplines, I cannot claim that the view of human deliberation that is based on the originator conception is anything more substantial than a view. It cannot be a robust foundation for conclusions about the true nature of the world. If we took it to be something more substantial than a view, and attributed reality to what was seen when we adopted the originator conception, we would make unjustified claims. If, for example, we were to take the supposed extraneous interventions to be real events in some non-physical realm, we would advance a claim that was at best unverifiable, and at worst a mere indulgence in mysticism.

I limit the degree of mystery that inevitably attaches to use of the concept of subject origination by accepting that the view that we get is only a view, and nothing more, although I would need a generous notion of verification, one that was not limited to reliance on the physically measurable, in order to rebut a charge of creating space for unverifiable claims about instances of human deliberation. The challenge is to show that this minimal position allows the view to have sufficient status, both to support our self-conception and to accommodate the phenomenology of deliberation.

A rather stronger claim than that of subject origination, a claim that agents are causes in a way that cannot be reduced to the causation of events

by other events, is the core of the various doctrines that go under the name of agent causation. The claim is stronger because agent causation is generally taken to be as real as event causation, although event causation itself may or may not be taken to be real in any strong sense. Subject origination as advocated here might be construed as a variant of agent causation without this presumption of equal reality, although the accuracy of that construal would depend on its precise content. My claim is not that we are points of origin, but that it is legitimate and productive to see us as points of origin. I do not formulate the doctrine of subject origination as a variant doctrine of agent causation, for three reasons. First, it is more straightforward to define a position from scratch than it would be to identify a specific form of agent causation from among those on offer, modify it and then add enough commentary to ensure that the desired position was stated. Second, I am concerned with processes of explicit deliberation that lead up to choices of action or adoptions of belief, whereas the proponents of agent causation are mostly interested in actions in general, whether or not they follow explicit deliberation, and are often not much interested in adoptions of belief. Third, I wish to avoid commitment to the libertarian incompatibilism that often, although not always, accompanies the advocacy of agent causation. In broad terms, I would be an incompatibilist if I did not think that we could make do with merely seeing ourselves as if we were points of origin. I shall return to agent causation in sections 3.7 and 3.8.

3.2 Phenomenology

In section 1.4, I set out an argument from phenomenology to the desirability of seeing ourselves as deliberating with mastery. I shall now set out an additional argument from phenomenology, not to mastery but directly to the originator conception. It is the conception that fits our experience of life from inside our own heads, and in particular our experience of deciding what to do and then acting.

When an agent deliberates and decides what to do, he has the clear impression that he is at the head of a causal chain, rather than being an

intermediate link in a causal chain. To the extent that the agent is aware of the influence of the past, he sees that influence as being in one of two forms, neither of which is seen as lying on a purely causal path to his decision. The first form is that of giving rise to facts that can be taken to be reasons for making certain choices, facts the status of which as reasons can be accepted or rejected, and the use of which in deliberation is under the agent's control. The second form is that of causes of the current physical state of the world, a state that will open up some options and close off others. Influences of this second form are causes of the range of options for the substantive decision as to what to do, not causes of the choice that is made. There is substantial overlap between these two forms. The physical state of the world can be captured in facts, some of which can be taken to be reasons for choosing certain actions. But this overlap does not amount to extensional coincidence. Some facts about what happened or about what was done may be taken to be reasons for making certain choices, without those facts being reflected in the current physical state of the world.

Adopting the originator conception allows us to do justice to this impression that we are at the heads of causal chains. It allows us to see ourselves as we appear to ourselves, as people who consider what to do, choose options and act, with our choices not being determined by natural antecedents. Even if someone making a decision believes in general terms that the choice he is about to make will be determined by antecedent states of himself and of the rest of the world, and by the operation of natural laws, because he knows that his mental processes are closely tied to physical processes, he still has a sense that the path to the outcome runs through a stage at which he weighs considerations and constructs an argument, from a position of mastery, rather than running only through purely causal chains, and that he is to that extent at the head of a new causal chain. Under my model of deliberation, the free choices are seen as made at the selection stage, but the consequence is that the substantive decision is also seen as originating with the subject. A given substantive decision is reached because of the choices that are made at the selection stage.

This argument is an argument for adopting the originator conception in general, not specifically in relation to deliberation. The strongest sense that we have of being at the heads of causal chains is associated with our

actions, rather than with our deliberations. The agent has a sense of causing something to happen, while his action appears to him to be uncaused. But if we adopt the originator conception in order to accommodate that phenomenology, the conception is then available for use in relation to deliberation.

It might seem that we should not allow phenomenology to have much influence on our consideration of philosophical questions. The natural sciences give us ample grounds to think that the world works in ways that are quite different from the ways that surface appearances would suggest. But I see no reason not to allow phenomenology a role, given that I seek a philosophy that will be adequate to our experience, and that I am only using phenomenology to argue that we should see ourselves as points of origin, not to argue that we are in fact points of origin. I would not support reliance on phenomenology to tell us facts about the world.

3.3 Entering into other people's heads

In this section, I shall discuss the notion of entering into other people's heads, treating their experience as if it were our own. We need to be able to do that in order to relate to people as subjects rather than as objects. I shall only discuss our relationship to others as deliberators. I shall not discuss the totality of our social relationships. Having argued for the need to enter into other people's heads, I shall elaborate on what is involved in entering into someone's head for this limited purpose. The notion of entering into someone's head will turn out to be a useful tool in determining the extent and the limits of sensible attributions of subject origination.

I refer to entering into heads rather than to standing in shoes, in order to indicate something more intimate than imagining oneself being in the same external circumstances as another person. "Putting oneself in another's shoes" can imply something more intimate than that, but it need not. The disadvantage of referring to heads rather than shoes is that it suggests a picture of a Cartesian soul moving from one head to another, something that would not be suggested by a picture of a whole body

stepping into a pair of shoes. No such Cartesian implication is intended.

In relation to experiences of deliberation, an observer's being able to treat a deliberator's experience as if it were her own would amount to her being able to have gone through the same sequence of thoughts as the deliberator, and to have experienced that sequence as a process of deliberation. Someone who could not do that might say that she could not have gone through the sequence of thoughts at all, or she might say that she would have had a cascade of thoughts that she could not have experienced as a rational interior monologue that had the form of deliberation. (The latter option would not amount to being able to treat the deliberator's experience as if it were her own. She must be able to treat the experience as if it were her own, without drastically altering the place of the process in the economy of thought. Apart from anything else, it would not be the same experience, except perhaps at the most atomistic level of qualia, if it were not experienced as deliberation.) The requirement to be able to have experienced the same sequence of thoughts as a process of deliberation does not directly import any requirement for the observer and the deliberator to have qualitatively similar inner experience. But there may be an indirect requirement when emotions play a significant role in deliberations, or when the contents of thoughts cannot be divorced from their occurrence in particular types of mental life. Such an indirect requirement could arise in connection with ethical or aesthetic thoughts, but it would not arise in connection with mathematical thoughts.

Why should we need to be able to treat other people's experience of deliberation as if it were our own? The reason is that this is the only way for us to see it as experience at all. We need to be able to do that because without that ability, we would not be able to relate to others as people who were like us, as members of the same society who chose actions and adopted beliefs in the same ways that we did.

There are two reasons why seeing experience as experience requires us to be able to treat it as if it were our own experience. One reason is related to experience in general, and the other is specific to the experience of deliberation. The general reason is that experience has to be appreciated directly. It cannot be seen as something that goes on in the world and that we can only observe, on pain of losing its character as experience. If

someone describes to us experiences of which we have no direct inkling, we cannot get any sense of his experiences as such. The reason that is specific to the experience of deliberation is given by the indexicality argument that is set out by Karsten Stueber, although he makes a different use of the argument (*Rediscovering Empathy*, pages 161-165). The argument is that we can only see thoughts as reasons for actions if we see them as thoughts that have indexical content for the thinker, rather than as thoughts that happen to occur to a person. The only way for someone other than the thinker to get that indexical content into her grasp of the thoughts is for her to see the thoughts as ones that might have been her own thoughts. There is a third reason, given in the words of Immanuel Kant: "If one wants to represent a thinking being to oneself, one must put oneself in his place, and so foist one's own subject on the object that one wishes to consider" (*Critique of Pure Reason*, A353). This remark is easy to interpret in a way that makes it obviously correct if one does not attach highly specific senses to the notions of representation and of putting oneself in the other's place. It would be harder to find an interpretation that was uncontentiously correct if one did attach highly specific senses to those notions, for example, senses that would make Kant an advocate of simulationism avant la lettre. I shall, therefore, not rely on Kant's point.

What is involved in entering into someone's head?

The notion of entering into another person's head is, like the related notions of standing in his shoes and of empathy, clear enough so long as one does not demand great precision. But it would be good to have as much precision as might be had without unduly restricting the notion, and without forcing commitments to specific philosophical positions that one did not have independent reason to adopt. There are after all many different, but related, notions in play in the general area of entering into someone's head, as has been demonstrated in a different context by Peter Goldie (*The Emotions*, chapter 7).

In the context of this book, there is no need for a notion of entering into people's heads that would cover all experiences. The focus is on deliberation. There is no need to cover the sharing of someone's emotions, or sense of excitement, or anything like that, except to the extent that the

impact of emotions and the like must be felt in order to have a direct appreciation of an instance of deliberation. That extent may be modest, but it is not negligible. A subject may, for example, choose an adventure after due deliberation because he feels excited about it, and the act of choosing may itself feel exciting. An observer could not appreciate what it was for the subject to go through that process of deliberation, allowing her to treat the experience as if it were her own experience, without having a sense of what it was to be excited at the prospect of an adventure. The process of making the choice would be so greatly influenced by a desire for excitement that the observer could not do justice to the nature of the process, unless she had a direct appreciation of what it was to be excited. Without that direct appreciation, she would be unable to follow the sequence of thoughts as a process of deliberation, because she would not grasp how excitement could be the driving force of the train of thought while that train remained a rational and controlled process. Only an observer who recognized the feeling itself would find that a description of the feeling was enough to make sufficient sense of the process as one of deliberation.

Once we have narrowed the field of interest to that of deliberation, the first point to note is that entering into the heads of others need not be a practice that is devoid of standards, a mere general feeling of empathy. There is no need for a practice that has something to do with empathy to lack standards. Consider for a moment other practices that also have something to do with empathy. We do not see a lack of standards in the use of the methods of Verstehen, understanding, and of Nacherleben, re-experiencing, that were identified by Wilhelm Dilthey ("Das Verstehen anderer Personen und ihrer Lebensäußerungen"). Max Weber's methodological remarks make the point more strongly (*Wirtschaft und Gesellschaft*, part 1, chapter 1, section 1.1, "Methodische Grundlagen"). He ties Verstehen as closely as possible to the use of public data. Indeed, he explicitly denies that Nacherlebbarkeit, the possibility of re-experiencing, is necessary, while acknowledging that it is useful (ibid., paragraph 2). He does not discard empathetic understanding, but shows that its use can be controlled by embedding that use in a procedure that is not itself empathetic. We can also note the rigorous analyses that are used

in social anthropology, where empathy is involved in the sense that the fact that anthropologists are human determines their approaches to the societies that they study. (For examples that are given in the context of a survey of the whole discipline, see James, *The Ceremonial Animal*, pages 56-59 on classifications as lived-in models of the world, and pages 79-83 on dance and its social and political roles.) The possibility of rigour in empathy is important because we should not distance entering into someone's head from empathy, in the way that we might be tempted to do if we equated the notion of entering into someone's head with some version of simulationism. Simulationism is a theory of how we manage to work and play together, and of how we sometimes manage to predict one another's conduct. What is important for success is that the simulator should start from the same position as the person simulated, and should end up with thoughts close to those that the person simulated ends up by having. Empathy is not intrinsic to simulation, nor need it be invoked in order to distinguish simulationism from theory theory. My concern here is with the recognition that another person's process of deliberation is like our own, not in the sense that it has the same qualitative feel but in the sense that we could have experienced the same sequence of thoughts as a process of deliberation. This is different from the concerns that simulationism is intended to address. Empathy must be involved. But it is not a mere general feeling. If we take entering into a deliberator's head to involve treating his sequence of thoughts as if it were our own, that picks out a particular form of empathy. It amounts to thinking on the same wavelength, rather than having the same inner sensations.

A consequence of the definition is that there are identifiable limits to the range of heads into which we can enter. Entering into the head of someone who is engaged in deliberation requires seeing his sequence of thoughts as something that one could have experienced as one's own process of deliberation. There is no need to require the observer to use a particular mechanism, such as running a simulation of the subject's mind inside her own mind, or engaging in something along the lines of method acting in order to forget herself and immerse herself in the character of the subject. (Some possible mechanisms are, for a different purpose, analysed in Goldman, *Simulating Minds*.) Nor need we require the observer to assume herself to

retain her own psychological make-up, values and knowledge. She should allow herself to imagine coming to resemble the subject to whatever extent was necessary, except that she should be limited to hypothetical adoptions of the subject's make-up, values and knowledge that would be within her actual capacity. This constraint is needed in order to avoid vacuity. Any observer could always say that if she were like the subject, she would have experienced the subject's sequence of thoughts as a process of deliberation. But it would not always be possible for her to say that if she were as like the subject as she could imagine her actual self becoming, she would have experienced the subject's sequence of thoughts as a process of deliberation.

We can identify some of the obstacles to entering into another person's head that might arise. One obstacle would be a bizarre logic, such as one in which modus ponens did not operate. Another obstacle would be a set of values that the observer could not imagine herself holding, such as a set that led to extreme human suffering's being accorded minimal ethical significance. A third obstacle, which might overlap both with the first and with the second, would be a bizarre standard of relevance of contextual features. Context matters in practical reasoning, making some chains of reasoning go through and blocking others. (For an analysis of the phenomenon see Price, *Contextuality in Practical Reason*.) We would find it hard to imagine not considering certain features of a context to be of great significance. We would, for example, expect someone who was thinking about whether to grow a new genetically modified crop outside the laboratory for the first time, to attach great significance to the scope for the crop to reproduce in an uncontrolled fashion. It would be hard to imagine anyone not taking much trouble to assess that risk. We would also find it hard to imagine a substantial lack of coherence between our background beliefs and our evaluation of the significance of consequences of given actions. We would, for example, find it hard to enter into the head of someone who was a great advocate of freedom of speech in general, but who thought that when debating the merits of proposed laws to ban speech that might offend members of particular religious or ethnic groups, the consequences for freedom of speech should not be given significant weight.

Attributions of subject origination can be regulated by virtue of such limits. If we could not experience a sequence of thoughts as a process of

deliberation, then we would be prevented from seeing it as an instance of deliberation with mastery. That in turn would eliminate any motive to attribute subject origination. We would not be completely prevented from seeing deliberation with mastery, but we would be effectively prevented because seeing deliberation with mastery would be entirely unmotivated, unless we grasped the process as something that we could have experienced as a process of deliberation. Without that grasp, any attribution of deliberation with mastery would be contentless. It would therefore be unmotivated. The notion of deliberation with mastery is one that is only given content by the experience of deliberation. There are no natural facts about the world that correspond to mastery. I set out an additional control over attributions of subject origination in section 3.6.

The obstacles to entering into another person's head would be obstacles to seeing someone as deliberating in a way in which we might ourselves deliberate. They would not be decisive obstacles to understanding in a detached way that someone's brain had worked in a given way to yield a given conclusion. To that extent, we could still attribute deliberation, even though any attribution of deliberation with mastery would be empty and unmotivated. We would simply regard the subject as rather strange. Any civilized society would, however, simply ignore the actions and beliefs of a human being who was like that, and would give him the space to do his own thing, so long as he did no harm to others.

Finally, three caveats must be entered. The first caveat is that claiming that one would be able to have a given experience is not the same as actually being able to have it. Such claims should, however, be a good enough guide for our purposes. They should be good enough because of the (usually) fairly dispassionate nature of the experience of something as a process of deliberation. Thinking about having experiences of this particular type is not likely to be significantly easier or harder than having them. But this does not hold when strong emotions or other feelings play conspicuous roles in a process of deliberation. Faced with examples of that nature, we would have to proceed more cautiously. We would need to be more ready to doubt an observer's claim to be able to have the required experiences, especially if the emotions or other feelings were not commonplace ones.

The second caveat is that there is unlikely to be any independent check

on an observer's claim to have a direct experiential grasp of a sequence of thoughts as a process of deliberation. We cannot yet measure the states of neurons and read off experiences to any more than a very limited extent, nor can we predict with confidence how sophisticated such technology might eventually become. We can rely on the fact that we live in a society in which what people say makes sense to other people, and we can reasonably hope that our interactions would detect at least some groundless claims to have direct experiential grasp, but that is all.

The third caveat is that we must work with a broad concept of experience, as set out in section 1.1. If someone deliberates and decides, she does not experience the process in the same way that she experiences the world through her senses. There is nonetheless an experience of deliberation and of decision. Those who disagree with this, and who think that there is nothing apart from passive experience of the world (including bodily states) and the non-experiential recognition of facts, such as the fact that one has made some decision, will be unable to accept the argument.

3.4 The range of attribution of subject origination

One consequence of the fact that use of the concept of subject origination is motivated by our self-conception and by the phenomenology of deliberation, is that the concept is unlikely to find much application in relation to non-human entities. There are two aspects to this. One is the lack of motive to attribute subject origination, and the other is the difficulty of attribution even when we have a motive.

Lack of motive

If we consider the computers and robots that we have made or that we can foresee ourselves making, we strongly suspect that they have no inner experience. In particular, we do not think that they have any sense of being points of origin. We can imagine what inner experience they might have if they had any at all, because their outer lives are often small fragments of what could be the outer lives of human beings who were enhanced with fast processor chips, infallible memories or strong arms, and because we

arrange for them to receive inputs from the external world that are often analogous to the inputs that our sense-organs allow us to receive. These things follow from the fact that we design computers and robots so that they can do specific jobs that arise in the context of our own way of life. If we did ever come to believe that they had inner experience that was like our own, we might then think that they would have a sense of being points of origin, so that we would be tempted to attribute subject origination to them. But even then, there might be no point in attributing deliberation with mastery to them, because we might feel no need to regard them as our equals or to integrate them into our society. We would also be strongly deterred from attributing deliberation with mastery or subject origination to them by the fact that their workings were obvious to us, an obviousness that would follow from the fact that we had built them. It would be perfectly clear to us precisely how they fitted into the causal network. It might be thought that this last point would extend to human beings. It is true that the more we come to know about how our brains work, the harder it will be to view ourselves in ways that imply any form of independence from the causal network. But the positive reasons that we have for attributing deliberation with mastery and therefore subject origination to ourselves, our desires to support our self-conception and to accommodate the phenomenology of deliberation, will continue to have force.

Difficulty of attribution

If we were to encounter aliens, we might well want to attribute deliberation with mastery and subject origination to them. If it was evident from what they had achieved that they were as sophisticated as we were, then we would seek to understand them as beings who were like us, at least to the extent of seeing them as beings who used evidence in deliberations that led to choices of action and to adoptions of belief. But we might find that it was very difficult to interpret the conduct of aliens in anything like the terms in which we interpreted human conduct, even at the minimal level of attributing deliberation, with or without mastery. The attribution of deliberation to aliens would be particularly difficult when their lifestyles were so different from ours that we could not even interpret their actions sufficiently to relate those actions to pieces of evidence that we might have

hoped to see them as considering to be reasons. In such circumstances, we could only attribute an empty form of deliberation. We would not be able to attribute content to the aliens' deliberations. Such an attribution would be comparable to the attitude of an observer who saw someone's actions as rational by standards that were incomprehensible to her. I shall argue in section 4.2 that such an attitude would make no sense.

Sometimes aliens would be closer to us than that, and we would have no difficulty in seeing them as deliberating. Could we then attribute deliberation with mastery and subject origination to them? We might very well want to do so, if we regarded them as rational beings who were like ourselves, and if we were able to enter into social relationships with them. They in turn might want to attribute subject origination, or something that was analogous to it, to us. But if the aliens were very different from ourselves in their inner experience, we would not be able to attribute subject origination to them in any contentful way. Our lack of grasp of the aliens' inner experience would include a lack of grasp of anything that might correspond to our sense of being points of origin. And the application of our concept to them would then misrepresent their inner experience.

3.5 The sufficiency of subject origination

I have argued that we should see ourselves as points of origin in order to support our self-conception, and in order to accommodate the phenomenology of deliberation. I now need to argue that if we merely see ourselves as points of origin, and overlook the causal closure of the physical, that is sufficient to do this work. There are two aspects to the question of sufficiency. The first is that of sufficiency of content. The second is that of sufficiency of status. Can a mere way of looking at our processes of deliberation do the work?

Sufficiency of content

The concept of subject origination does not have enough content to do its work in isolation. I need to show that the necessary content can be supplied

from other sources, in a way that is consistent with the content of the concept of subject origination.

The content that subject origination affords us has already been described. The concept allows us to see ourselves as having mastery over the choices that we make at the selection stages of deliberations, choices that are seen both as open and as under our control. I shall identify three deficiencies of content, and indicate how the deficiencies might be supplied. The point of doing this is to show that there is no danger of setting up a concept of subject origination that is sufficient to do the very limited amount of work that it can do on its own, but that is left isolated from the resources that are needed for it to be of any greater use.

The first deficiency of content is that use of the concept of subject origination does not in itself give us access to inclinations, emotions or other aspects of our mental make-up. It is vital that the wider content of our minds should be available. Our concept of deliberation does not amount to the selection of one from a range of options that are written neutrally in black ink on a white sheet of paper. Every option is already coloured by our inclinations and by our emotions. A concept of a process of deliberation as abstract and colourless would not be our concept at all. It would reduce decision to random selection. The fact that our deliberation is not wholly dispassionate is one truth in David Hume's remark, "Reason is, and ought only to be the slave of the passions" (*A Treatise of Human Nature*, book 2, part 3, section 3, page 415). The mastery that we can see ourselves as having contradicts Hume's imputation of slavery, if we take the passions to be among the rational antecedents. But we do see ourselves as masters over ourselves, with all of our characteristics, not as masters over mere ciphers. Fortunately, use of the concept of subject origination does not rule out access to the wider content of our minds. Use of the concept only prevents us from seeing that wider content as all-powerful.

The second deficiency of content is that use of the concept of subject origination does not in itself give us access to descriptions of external circumstances. Such descriptions, including the histories that lead up to our choices, must be included in order to give a full picture of our deliberations, and they are not included merely by virtue of seeing someone as a point of

origin. They are supplied by records of the facts, as they appear to the subject, as they appear to any other people who are involved and as they might appear to disinterested observers. Such records are available on the basis of the naturalistic conception of ourselves and of the world, supplemented by the originator conception that accommodates our understanding, from the inside, of what it is like to weigh evidence. Use of the concept of subject origination does not cut us off from the external facts.

The third deficiency of content is that use of the concept of subject origination does not in itself give us any way to discriminate between the rational and the irrational. Seeing oneself as a point of origin does not make one's choices of action or adoptions of belief rational. We do need standards of rationality, and we cannot leave them entirely up to the individual. I shall address this point by arguing in sections 4.2 and 5.3 that rationality can be taken to be rationality for us, with the standards of rationality coming from our society.

Sufficiency of status

I now turn to the claim of sufficiency of status, the claim that the concept of subject origination can do its job even though it is only to be taken to give us a way of looking at our processes of deliberation. Its use should not be taken to imply the identification of anything in the world beyond the things that would feature in a naturalistic inventory, nor to imply the attribution of additional real properties to those things.

One could dismiss the concern by pointing out that the job is to support a mere self-conception and to accommodate a mere way in which things seem, the phenomenology of deliberation. A mere way of looking at our deliberations should suffice to support conceptions and to accommodate appearances. But that would be too quick. It would not take account of the possibility of an argument that paralleled the no miracles argument in the philosophy of science. I shall turn to that argument first, and then test the claim of sufficiency by reference to our need both to explain what we do and to attribute responsibility for our actions.

No miracles

The concept of subject origination is remarkably useful. Not only do we get

a view of humanity that supports our self-conception. We can also use that view to give coherent and satisfying accounts of our own and other people's lives. Such accounts can be given when we see people as making free but controlled choices. A satisfying account of a human life is an account that mentions such choices, rather than an account that is given wholly in terms of mechanical responses to the world. Biographers may try to explain why their subjects chose certain actions and adopted certain beliefs, and we do understand people better, and find accounts of their lives more satisfying, when we are shown the long-term influences of their education or the momentary pressures under which they made important decisions. But if a biographer succeeded totally, so that a life was seen as the inevitable working out of impersonal historical forces, whether long-term or transient, global or personal, the subject would be reduced to a cipher and the account would not satisfy us. (In order for the concept of subject origination to make its full contribution to the formulation of coherent and satisfying accounts of human lives, we need to extend its use beyond choices where systematic reflection is expected. When we use the concept in relation to simpler choices, it may be hard to discern a selection stage and then a calculation stage. But the extension to simpler choices is perfectly legitimate. If all that we can see is that someone simply acted on reasons, we can say that there was a selection stage at which the reasons in question were automatically given substantial weight and at which some obvious method of argument was chosen by default. My model of deliberation can be applied because when a choice is made for reasons, the subject must accept that the reasons are significant enough and must make use of them in some specific way. The subject could have attached no weight to the reasons, or could have used them differently, even if he did not in fact pause to consider those options.)

It is odd that the concept of subject origination is so useful. A leading argument for scientific realism is the no miracles argument, as set out by Hilary Putnam ("What is mathematical truth?", page 73). The claim is that the success of the natural sciences can only be explained if scientific theories state how the world actually is, given that otherwise their success would be a miracle. The success of the argument in establishing scientific realism is contested, and I make no claim that it succeeds. But the argument

clearly needs to be answered, and a parallel argument could be formulated for subject origination. Given that it is so productive to see people as points of origin, allowing us to see them as deliberators with mastery, it is odd that use of the view of someone as a point of origin does not involve a claim about the true nature of the world. Furthermore, it cannot involve such a claim, on pain of metaphysical implausibility.

In fact, the issue is independent of the debate between scientific realism and anti-realism. I deny that a process's being seen as one of deliberation with mastery, and hence as one in which the subject is a point of origin, is a matter of its having some complex of natural properties, even though each occurrent thought must be embodied in a physical entity. The debate between realism and anti-realism arises in connection with entities and properties that are accepted, on all sides, to be the concern of the natural sciences. Once we move away from natural entities and properties, as we do when we consider subject origination, we must reconsider the validity of the claim that it would be a miracle if successful explanations did not refer to real things or their real properties. The significance of moving away from the natural realm is that we do not then seek to explain what happens in the way that scientific accounts explain what happens. We do not seek contrastive explanations, a point to which I shall return later in this section. Arguments that are used in the debate as to whether realism or anti-realism is the appropriate philosophy for the natural sciences cannot then have any more than analogical relevance.

A related consideration weakens even the analogical relevance of arguments that are used in the debate in the philosophy of the natural sciences. A mainspring of that debate is that there are, or have been, several theories covering the same ground, as with the Newtonian theory of space and time and the Einsteinian theory of space-time. We expect contests between theories to be resolved, unless it can be shown that the theories are equivalent. That is, we expect some naturalistic theories of the world to be better than others, and we expect to be able to make at least a provisional judgement as to which of several competing theories is the best available theory. When we see some of our mental processes as involving subject origination, we have no such ambition. We merely claim that processes of deliberation make sense to us if they are seen in the given terms. We do not

claim that this is a better way of seeing them than any other way. We do not, for example, assert that the originator conception should supplant the naturalistic conception.

Explanation and responsibility

We expect theories to explain what happens, including human choices of action and adoptions of belief, in at least some sense of explanation, although not necessarily the scientific sense. We also expect our conception of human action to give us grounds for attributing responsibility, so that we feel justified in apportioning praise and blame, reward and punishment. How does the concept of subject origination fare? On the one hand, application of the concept allows us to see people as standing back and freely considering the evidence and how to use it when choosing actions and adopting beliefs. An imputation of subject origination opens up a route to the imputation of a strong form of personal responsibility. On the other hand, the concept only gives us a way of looking at what people do, and we might expect that fact to limit what we could get out of its use.

On explanation, subject origination fares well so long as we only seek to make sense of what people do, and do not seek contrastive explanations or predictive powers. If we give an account of someone's conduct in terms of free but controlled choices, terms that imply subject origination, the account can make sense to us. We can get a coherent and satisfying account of the subject's life, an account which allows us to identify with the life as our kind of life. This is particularly so if we extend our use of the concept of subject origination to choices where systematic reflection is not expected, as described above. This is not a matter of seeing why the subject did certain things. We could see that merely by considering the external facts that the subject might take to be reasons for certain choices, along with facts about the prior states of the subject, and working out how a human being in those states, faced with those external facts, would be likely to respond. Rather, it is a matter of seeing a life as one that the subject leads, and not as one that leads the subject. A life that the subject leads is the sort of life with which we can identify, and the sort that makes sense to us. A way of looking at the subject's processes of deliberation should suffice for this purpose, because a mere presentation of a life should be

enough to allow us to identify with that life. Identification with a life does not require that the life should in fact be of the nature that is imputed to it, so long as one is faithful to the natural biographical facts. Such facts are not contradicted by viewing the subject as engaging in subject origination.

Accounts that make sense are important. Jaegwon Kim draws our attention to the need to be able to give accurate rational accounts of what we do, simply in order to keep a grip on our notion of agency ("Reasons and the First Person"). A mere way of looking at the subject's life should suffice for that purpose, although it is also true that an account which mentioned reasons without mentioning freedom, and which made no use of the notion of subject origination, would suffice for Kim's specific purpose. Moving up to the largest scale, Alasdair MacIntyre has argued for the importance of narratives that stretch through whole lives, and beyond them, in making sense of human actions (*After Virtue*, chapter 15). Furthermore, the whole-life narratives would need to be of types with which readers could identify. Actions only come to make sense to us, in the way that MacIntyre describes, when a life within which they are set is seen as our, human, kind of life, a life that involves free but controlled choices. For these purposes too, a way of looking at the subject's life should suffice.

What a mere way of looking at a process of deliberation cannot give us is a contrastive explanation, an explanation of why the subject chose one option rather than another (at the substantive level of the action chosen, rather than at the selection stage of deliberation). An ability to construct contrastive explanations would be tantamount to an ability to predict. A contrastive explanation would amount to saying that X rather than Y happened at a given time because conditions at some earlier time, or contemporaneous environmental conditions, were Z. That sort of grasp of the way in which the world worked would give the tools for prediction. We cannot often predict the course of a life as it progresses, but we can still see how the course that is taken makes sense. Sometimes we are not at all surprised at a subject's choices, because the reasons for making those choices rather than the obvious alternatives strike us as overwhelming. On other occasions, we are surprised by what someone does, yet in retrospect see the choices that he made as fitting into a coherent life, or a segment of a life, that he has led. If we can see that, then we have a form of

explanation. But as with whole-life narratives, the explanation will only carry conviction if it is given in terms with which we can identify. That will require it to refer to choices that we see as free but controlled. The fact that the lack of a contrastive explanation of the course of a life, and the associated lack of predictive ability, do not amount to a lack of explanation is one of the truths in Kierkegaard's comment, "It is quite true what philosophy says, that life must be understood backward. But then one forgets the other principle, that it must be *lived forward*" (*Kierkegaard's Journals and Notebooks*, volume 2, JJ:167, page 179; Danish text in *Søren Kierkegaards Skrifter*, volume 18, JJ:167, page 194).

The distinction between contrastive explanation and explanation does not depend on any thought that prediction would, impractically, require us to trace micro-physical causal chains. Nor does it depend on any thought that contrastive explanations must be couched in physical terms. Suppose that we confined ourselves to the concepts and propositions that were used by the subject, with no reference to states of his brain, and went on to make two claims. The first claim would be that a subject would always perform a given action if the reasons normatively required that action and if there was no specific obstacle to the action. The second claim would be that given the first claim, prediction was possible, even within the confines of the concepts and propositions that were used by the subject. Even then, the availability of explanations would not imply the availability of contrastive explanations. As Joshua Gert has argued, there is a gap between reasons that justify actions and reasons that require them, and the gap is not merely a matter of the quantity or weight of reasons (*Brute Rationality*, pages 20-21 and passim). We could have justifying reasons, explaining actions in retrospect, which did not require those actions. The justifying reasons would therefore not have made the actions predictable, even on the basis of the two claims that have just been proposed, and however weighty the justifying reasons might have been. It should, however, be noted that Gert explicitly limits his claim to the practical realm, and does not extend it to the theoretical realm of the adoption of beliefs. So his argument does not lead to doxastic voluntarism. Furthermore, his concern is with justification and requirement, notions that are applied to actions themselves, rather than with the process of deliberation. Thus his argument

does not yield the sense of control over the influence of reasons on our choices, at the time of making them, with which I am concerned. Another useful perspective on the issue is given by Carl Ginet, who argues for the sufficiency of anomic reasons-based explanations of actions ("Reasons Explanation of Action: An Incompatibilist Account").

Moral responsibility for actions would be very hard to establish on the basis of a mere way of looking at processes of deliberation. We can say that attributions of responsibility in themselves make perfect sense. Within a narrative that we construct about a person in which we apply the concept of subject origination, we say that certain actions were his actions, that they were chosen by him, so that he is responsible for them in the sense of owning them. But that leaves unanswered the question of whether we should move on to responsibility in the moral sense, the sense that legitimizes reward or punishment. To put the question in an acute form, how could we possibly think that a mere way of looking at people, a way that deliberately ignored the full effect of physical causation, was enough to justify putting people in prison?

There is a large and inconclusive literature on the relationship between different concepts of free will and the legitimacy of attributions of moral responsibility. Use of the concept of subject origination is not going to yield a solution. If anything, use of that concept would encourage the recognition of a fissure in our thinking. On the one hand, we may construct narratives that presuppose the legitimacy of seeing subject origination, and we may use those narratives when attributing responsibility, in the sense of attributing ownership of actions. (We might also make such attributions independently of narratives of that type, but without the benefit of undeniability of ownership that was noted in section 1.4.) On the other hand, we may reward and punish people in order to encourage conduct that we find especially desirable and deter or prevent conduct that we find especially undesirable. The attribution of ownership of actions is a necessary but insufficient conceptual support for the legitimacy, as opposed to the usefulness, of practices of reward and punishment. Grounds for a move from the attribution of ownership to the attribution of moral responsibility would make good the insufficiency in the support. The search for such grounds is the search for a way to get an ought from an is,

or rather, in the current context, the search for a way to get the ought of reward and punishment from the as-if-it-is of humanity as seen when we implicitly or explicitly attribute subject origination. Alternatively, one could re-think the conditions for attributing moral responsibility. One could, for example, argue that the agent's possession of guidance control was sufficient (Fischer and Ravizza, *Responsibility and Control*).

3.6 The legitimacy of subject origination

The argument so far has been that there are good reasons to attribute subject origination, and that the concept can do its job even though use of it only amounts to adopting a way of looking at our processes of deliberation, and does not imply a claim about the contents of the world. This is not enough to make a full case for use of the concept. We should also consider the legitimacy of our use of the concept, by reference to general standards. As part of that work, we need to consider the significance of the lack of scientific explanation for subject origination.

Subject origination and standards of legitimacy

I claim that it is legitimate to see our deliberations as conducted with mastery, and hence as involving subject origination, and to overlook the causal closure of the physical. I do not claim that this is the only legitimate view. A scientific view that sets us, and our deliberations, entirely within the causal network is equally legitimate.

For a view to be legitimate, it must not contradict other legitimate views. Its adoption must also allow us to see something that cannot be seen when we adopt those other views, or that cannot straightforwardly be seen when we adopt them, otherwise there would be no motive for its adoption. Finally, what is seen must not be arbitrary. There must be some things that it would be a mistake to see. If that were not so, but the view licensed a free-for-all, then adoption of the view would yield no worthwhile information.

A view of our deliberations as involving subject origination does not contradict the scientific view, because the former view can do its job while we regard it merely as a way of looking at our deliberations, and do not

regard its adoption as involving any claim that a deliberator could in fact choose any one of a range of options. Even so, there is an issue to resolve. If we view our deliberations as involving subject origination, that implies turning down in advance the offer of the natural sciences to give comprehensive accounts of the phenomena that are of interest. Given the power of the natural sciences to explain surface appearances by deep structures that relate diverse phenomena, and the impressive, if incomplete, achievements of the natural sciences in deriving vast chunks of our knowledge from physics, this attitude is in need of some defence. It would be perfectly reasonable to challenge the ambitions of the natural sciences by arguing that some things were beyond their powers of explanation, although it is unclear whether such arguments would succeed. The arbitrary exclusion of the natural sciences from certain areas of work would be much more dubious.

To defend the view that sees our deliberations as involving subject origination against this charge, we must recognize that it is a view that can only be taken on the basis of the originator conception of ourselves, whereas the scientific view of our deliberations is one that can only be taken on the basis of the naturalistic conception of ourselves. I avoid a charge of illegitimately refusing to allow the natural sciences free rein by positing two different conceptions of ourselves, conceptions that answer to different needs but the relationships between which we can legitimately discuss. We can for example conclude, as I have, that application of the originator conception involves one specific difference from application of the naturalistic conception. We overlook the causal closure of the physical. Everything else that we see when we apply the naturalistic conception, including our spatio-temporal location, our biology and our ability to cause events to occur by acting, we also see when we apply the originator conception.

There are clear reasons for the difference between the two conceptions. They answer to the needs of two different types of observer of humanity, the human participant in human society and the more detached, scientific observer. Our inner experience and our self-conception give us good reason to adopt the originator conception. But when we observe humanity in our capacity as natural scientists, such things are of

no interest to us. The point is reinforced by the consideration that any results in the natural sciences should be accessible to a far wider range of rational beings than human beings. Many such beings would not even be able to comprehend our inner experience or our self-conception. Thus in the role of practitioners of the natural sciences, we need to stick to the naturalistic conception of ourselves. Only a conception that eliminates the specifically human point of view, and that limits us to that which would be visible to a much wider range of observers than human beings, is fitted to the conduct of the natural sciences.

We must ask whether this distinction between the originator and the naturalistic conceptions is of the right kind to legitimize ignoring some of the results of the natural sciences when we adopt the originator conception. Anything that we thought about ourselves and other people that contradicted scientific results would certainly be unacceptable. To that extent, our scientific knowledge should remain with us at all times. But we do not need to contradict our science in order to see our processes of deliberation in a non-scientific light. All we need claim is that if we describe the decisions that are involved in the selection stage simply as taken by us, that is an adequate description for the purposes at hand. Those purposes are to support our self-conception, to accommodate the phenomenology of deliberation and to make sense of our lives. A self-conception and the accommodation of phenomenology can both be based on a partial view of human beings, one that leaves out the naturalistic detail that would bring the role of the causal network to the fore. And we make sense of our lives by giving coherent and satisfying accounts of what we do. Again, a partial view is enough. By way of analogy, it is perfectly legitimate to describe a natural system in terms of the transmission of information, for example, the information that is transmitted by a dancing bee as to the location of suitable plants, while leaving out physical details such as an account of how bees manage to fly. We can abstract the task of conveying the information from its physical realization, and see how the same task could have been accomplished using different physical means.

This analogy may seem to be inappropriate. While we can abstract information from its physical embodiment, we can always trace particular physical embodiments and relate them to the information. We may, for

example, show how there are enough atoms, in an adequately manipulable and readable form, to hold and transmit the information. Such a link with the physical is much less obvious when we come to something that is expressly not to be captured by a scientific view, choices at the selection stage of deliberation that are open but that are controlled by the subject. But the link to the physical can still be made. In the case of information, the correspondence of information with a physical realization is token-token. A particular piece of information is, on a particular occasion, held and transmitted in a particular physical form. Different physical forms could be used on different occasions, the only requirements being sufficient manipulable and readable complexity in the physical material, and a context that was sufficient to yield semantics for the actual physical form that was used. That context would itself be embodied in a physical form that could be different while still allowing the context to perform its role. Likewise, individual instances of deliberation must be physically embodied, and the material must have sufficient complexity, but only a token-token correspondence is needed. No particular embodiment is required, so no characteristics of any particular embodiment need be associated with any one instance of deliberation. This allows a perfectly good analogy with the holding and transmission of information. The analogy is only meant to make the point that one can freely abstract meaningful accounts, omitting physical detail. There is no intention to use the analogy, or the token-token nature of the correspondence, to argue that an instance of deliberation could break free from all physical embodiments, and thereby actually break free from the causal network, so that it could actually have involved different choices from those that were in fact made.

Even so, a feeling of unease may remain. A theory of information can sit very happily alongside a theory of matter, and different parts of a single well-integrated body of mathematical knowledge are used to formulate both theories. The two theories even come together in dramatic ways, as they do in theories that relate the capacity to hold information to spatial regions and their boundaries (Bekenstein, "Information in the Holographic Universe"). Similar things could have been said if I had chosen any other analogy that was based on two different parts of the natural sciences. We do not see anything like the same integration between the view of humanity

that is given by the naturalistic conception and the view that is given by the originator conception. We even know that we must not seek integration, because if we adopt the naturalistic conception we must accept the causal closure of the physical, whereas if we adopt the originator conception we must reject it. Any feeling of unease reflects a recognition that we cannot claim that the view of humanity that is given by the originator conception is simply a picture of how things really are. We can, on the other hand, reasonably make that claim on behalf of the view that is given by the naturalistic conception. Unfortunately, we have to step onto the thin ice of views that cannot be said simply to tell it like it is, if we are to have a satisfactory understanding of our lives. A continuing feeling of unease is no bad thing, if it reminds us not to read too much into the deliverances of such views.

One other possible clash between the originator conception and the natural sciences must be considered. It would not be acceptable to identify a homunculus or a Cartesian soul lurking anywhere within a human being. John McDowell writes of the danger of applying a narrow form of the naturalistic conception, one that finds no place for spontaneity of action. He argues that such a conception would tend to drive us to locate spontaneity either in a non-natural realm or in a specifically inner realm (*Mind and World*, pages 89-91). McDowell's response is a more inclusive naturalism, one that finds space for the realm of reasons as our second nature. My proposal means sticking with a narrower naturalism than that which might easily result from pursuing McDowell's proposal. But I am not driven to regard the spontaneity that appears under the originator conception as something non-natural, because I only offer the originator conception as a way of seeing ourselves. There is no claim that we are in fact as we seem to be when we apply the originator conception, whether as a matter of natural or of non-natural fact, so the non-natural does not enter the picture. I also avoid being driven to locate our spontaneity in some inner realm, because when we view an episode of deliberation as involving subject origination, that does not require us to move the episode from an outer realm to an inner one. No additional actual event is identified, so no location, inner or outer, need be found for one.

The second requirement to establish the legitimacy of the view of our

deliberations as involving subject origination is to show that the adoption of this view allows us to see something that cannot be seen if we adopt other legitimate views, or that cannot straightforwardly be seen if we adopt them. This has already been established. If we see our deliberations without using the concept of subject origination, we see ourselves merely as creatures of the causal network. That would leave us without any obvious way to support our self-conception, or to accommodate our inner experience of deliberation. There are, of course, plenty of compatibilist proposals that would allow us to see ourselves merely as creatures of the causal network without undue discomfort, but their acceptability is disputed. And as argued in section 1.4, such proposals would not secure the desired status of the subject as essential and as in charge.

The third requirement is to show that what is seen when we view our deliberations as involving subject origination is not arbitrary. There must be some things that it would be a mistake to see. Some accounts of how specific individuals deliberate must be rejected. There is a perfectly good control over what it is correct to see. It is that what we see must yield coherent and satisfying accounts of people's lives. The accounts that we construct implicitly attribute subject origination because they refer to subjects as deliberating and as deciding, in a way that we see as both free and controlled and that we do not see as amounting to following programs that have been foisted on the subjects. If some such accounts do not make sense, then the deliberations should not be seen in that light and subject origination should not be attributed, whether implicitly or explicitly.

This control is nowhere near as rigorous as the controls in the natural sciences over the conclusions that may be drawn from experimental data. The control also relates only to entire accounts of lives or of segments of lives, possibly short segments, that are given in human terms. Such an account must include not only all of the significant events in a given period and the subject's relationships with other people, but also emotions and impulses. The control could not operate in relation to the very partial account that would be given by cataloguing only episodes of deliberation, or by cataloguing all actions along with a commentary that referred only to the subject's conscious reasons for his actions. It is also only rarely a control that applies directly to an isolated attribution of subject

origination. There are few occasions on which we cannot reasonably see a process of deliberation as having involved subject origination. Those are the occasions on which we cannot enter into the heads of the deliberators. Much more commonly, the control will be a control over accounts of sequences of decisions and of actions, including the decisions and actions that we would not expect to be preceded by systematic reflection. Is the agent seen as making decisions that we can, when we take them together as a sequence in a single life, relate to character, to motives and to circumstances in a way that presents the subject as someone who makes free but explicable choices and who leads a coherent life rather than being led? Fortunately, that gives us an indirect control over the attribution of subject origination. If we cannot see a sequence of decisions as both free and making sense, we will not see the subject as leading our kind of life. We will then have no motive to attribute subject origination, even though we may recognize that the subject's inner experience might give him a motive to attribute it to himself.

The control can work because we explain what people do in ways that make sense to us, given our own experiences of living, of making choices and of acting. Some accounts make sense and some do not. We can see this by considering novels. Extraordinary things may happen to characters, or the characters may inhabit physical environments or societies that are very unlike our own, but we still have a sense of whether the descriptions of their motives, their reflections and their decisions make sense. In a good novel they do make sense, and we can relate to the characters. In a bad novel, they do not make sense. The characters are too robotic, or their decisions are too outlandish. We can also see the importance of our possession of the materials with which to construct accounts that make sense by noticing the disconcerting effect of not having those materials. The film *Cinq fois Deux*, directed by François Ozon, gives five scenes from a relationship in reverse order, from divorce to first encounter. Only by watching the whole film can one properly interpret the scenes that are presented first. As scenes that are shown later upset initial interpretations of scenes that are shown earlier, confusion mounts.

If the control were based purely on the experiences of an individual who was taking a view of people's lives that involved implicit attributions

of subject origination, it would not be much of a control. Some individuals are inclined to formulate eccentric accounts of other people's lives. Fortunately, there is a social dimension to the control. An account of someone's life in human terms, an account that implicitly attributes subject origination and that involves our seeing the people described as like us, is an account that we can share with others. If we converse with others, we soon find out whether we interpret people in the ways that others do. If we do not, then we should reflect on our own ways of looking at people. It is possible for one person to be correct and for everyone else to be mistaken, but it is unlikely when we are dealing with matters of such widespread human experience as daily life. One demonstration of the existence of shared standards is the fact that it is possible to systematize the ways in which we interpret narratives in order to bring out the relationships between actions and other actions, events and situations. (One such systematization is given in Kafalenos, *Narrative Causalities*.)

The practical effect of the control that what we see must yield coherent and satisfying accounts of people's lives depends on the nature of our own way of life, as human beings. A way of interpreting human lives can yield interpretations of individuals that make perfect sense to us, but that would make no sense to at least some aliens. Similarly, their modes of interpretation of the conduct of their own kinds might yield interpretations that made no sense to us. That is, there is no universal standard of correctness to be had. This reflects the fact that we do not, when we attribute subject origination, see some objective and very widely visible natural feature of the world. Instead, we make use of a way of looking at our lives that reflects our own experience of life from the inside. Only others with a similar experience of life can be expected to be able to share that way of looking at our lives.

Coherent and satisfying accounts of human lives have a lot to be said for them. They are essential to our senses of who we are. It is also possible to argue for the strong, and debatable, thesis that "a life as led is inseparable from a life as told" (Bruner, "Life as Narrative", page 708). But two notes of caution must be sounded. The first cautionary note is that that there are philosophical arguments on the other side. Galen Strawson argues that consciousness of a personal narrative that unifies one's life in even a

moderately strong sense need not be important ("Against Narrativity"). The second cautionary note is that a desire to construct coherent and satisfying accounts can lead us astray. Take what happens when people are told about a woman who has studied philosophy, who is very concerned about discrimination and social justice, and who has been politically active. People often estimate that the probability that she is both a bank teller and an active feminist is higher than the probability that she is a bank teller. Including feminism makes for a more satisfying account, but the probability of a person's having two given qualities cannot be higher than the probability of that person's having a specified one of the two (Tversky and Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment", page 297).

This does not mean that the addition of beliefs to enhance coherence always diminishes probability. The reason is that probabilities that are conditional on evidence can have a role, as in the example of Tweety the penguin (Olsson, "The Place of Coherence in Epistemology", section 3). Olsson's paper as a whole examines the relationship between coherence and truth. His gloomy remarks in sections 7 and 8 about the failure of coherence to be truth-conducive do not worry me. I maintain that we seek coherent and satisfying accounts in order to make sense of human lives, regardless of whether that search improves the ratio of the number of our true beliefs to the number of our false beliefs.

The lack of scientific explanation

The acceptance of subject origination requires acceptance of a mystery, something that is not scientifically explained. I shall now describe this mystery, and then reflect on whether we can live with it.

The mystery is that of the person who is a point of origin. We have got used to scientific explanations of what lies beneath the surface of everyday objects, the molecular structures that explain their properties. We can go deeper, explaining molecules in terms of atoms and atoms in terms of fundamental particles. Wherever our science has got to, there is some bedrock, something unexplained that we just have to take as given. But that bedrock is far beneath our everyday world, and to that extent our everyday world is thoroughly explained. Furthermore, we may well find that today's

bedrock can be penetrated tomorrow, when we discover a new and deeper explanation that will become, at least for a while, the new bedrock. The person as a point of origin, on the other hand, is simply a given. We have a sense that we are such people, and the natural sciences may well come to offer us explanations of why we have that sense. But the natural sciences are not going to offer us an explanation of how it is that we are points of origin. Any naturalistic explanation would take us back wholly into the causal network, where no such points of origin could be seen.

This does not mean that there is nothing to be said to enlarge our understanding of the person as a point of origin. But in order not to destroy the supposed status of the person as a point of origin, anything that was to be said could not resemble a scientific explanation. At least three approaches are available. The first approach is poetic or other artistic insight into what it is to be human. The second approach relies on grasping other people's inner experience as experience, using the analogy of our own inner experience. Our own sense of what it is like to deliberate and to decide is a rich and familiar one, even though its detail may be hard or impossible to put into words that would be comprehensible to non-human beings, and we can take it that other people have similar senses of deliberation and of decision. The third approach is to note that if we take ourselves and others to be points of origin, we can understand both our own deliberations and the deliberations of other people by seeing that they mesh together, in a way that makes sense of both. The meshing together is evident when we make choices to which others respond, and when we respond to their responses and to their own choices. If we see everyone as deliberating in the same way, making free but controlled choices, we can gain insights into how societies function.

All three approaches are valuable, but none of them gives us explanations in a sense that would be comparable to scientific explanation. This is obvious with the first, poetic or other artistic approach. But all three approaches take a lot of what is to be explained for granted. The first approach takes for granted our ways of responding to poetry or to the other arts. That is, it takes for granted that we find such things meaningful, but we only find them meaningful because we respond as human beings who have a human appreciation of life from the inside. The second approach

takes a lot for granted more directly and more obviously. It works outward from our own experience, in which we feel that we are points of origin, using that experience as a surrogate for an explanation of what it would be for someone to be a point of origin. The third approach takes for granted that the problem of how to comprehend life as experienced is a problem to be solved, rather than one to be dissolved as a phantom problem. Only on that basis can the third approach be used to support the idea that we and other people are points of origin. The support is that seeing people as points of origin helps to explain the social phenomena of debate, of agreement and of compromise. It is taken for granted that these phenomena need an explanation in terms that are faithful to their appearance to us, as opposed to an explanation in the very different terms of brain cells and of human animals as they are seen biologically.

We are therefore left with the person as a point of origin, as something that defies scientific explanation or anything like it. Not only do we not have a scientific explanation. We can see that we are not going to get one, because a scientific explanation would inevitably locate the person wholly within the causal network, making it impossible to see the person as a point of origin. Is this a mystery with which we can live? I think that we can. The inexplicable is not intrinsically offensive, so long as it does not pretend to be of the same kind as something explicable. If we were to accept such pretenders, we would be in danger of wandering into the realm of purported supernatural forces. Encroachment on the province of physics, by anything other than that which is within the scope of current physical theories or that which we can reasonably hope to bring within the scope of future physical theories, is not welcome. We can avoid such difficulties by not taking subject origination to be a feature of reality. I shall return to the element of mystery and to the relationship with the natural sciences in section 3.8, in connection with a concept of the subject that allows us to see subject origination.

3.7 Other approaches

In this section, I shall discuss two approaches that are in play in the same

field as subject origination. The first is agent causation. The second is the theory of the intentional stance, as presented by Daniel Dennett. I shall end the section with comments on the idea of a psychological theory that is detached in the sense of not reflecting our inner experience. The approaches that I discuss in this section are of wide application. They can be used far beyond the narrow field of those choices of action and adoptions of belief where systematic reflection is expected.

Agent causation

Several positions that differ in detail can be gathered together as variant doctrines of agent causation. The core of these positions is that agents are causes in some special way that is not reducible to ordinary event causation. Doctrines of agent causation are controversial, and there are plenty of objections to them. Can the doctrine of subject origination avoid the objections?

It can avoid them, because the objections to doctrines of agent causation rely for their force on the claims of those doctrines to identify something that is in the world, just like event causation, and that explains actions by virtue of actually having causal efficacy. Subject origination, on the other hand, is specifically not claimed to be as real as event causation. Some examples of objections to agent causation will illustrate the point. (Discussions of objections such as the following include Clarke, "Agent Causation and the Problem of Luck"; Lemos, "Flanagan and Cartesian free will: a defense of agent causation"; O'Connor, "Agent Causation"; Schlosser, "Agent-causation and agential control"; Schlosser, *The Metaphysics of Agency*, chapter 1.)

Against agent causation, there is the argument that agents would need to be special substances that could be causes in themselves, unlike the everyday natural substances that are widely thought to enter into causal processes only through events that involve them. There is the objection that the notion of causation plays no explanatory role at the beginning of a causal chain, because to say that the agent caused the first event would only add something to saying that he was its agent if he did something to cause that first event, which would mean that it was not the first event (Davidson, "Agency", pages 52-53). There is the argument that only events, which have

the property of occurring at particular times, could explain why other events, including actions, should occur at particular times. There is the disturbing contrast between the ready application of the concept of causal sufficiency to the ordinary causation of events by other events, and the difficulty of applying that concept to agent causation. There is the related difficulty of attributing adequate control to the agent when the agent could have performed any one of a range of different actions. The notion of responsibility for one's actions, whether moral responsibility or more general causal responsibility, can be undermined, because it can seem to be a matter of luck which action an agent chooses. Finally, it can be argued that naturalistic causal accounts that involve events can relate our conduct to the reasons that we have for acting in certain ways, while accounts that invoke agent causation cannot do so because states of our brains that correspond to reasons could not have appropriate effects, given the assumptions that are needed in order to make sense of agent causation.

There are answers to all of these objections, and many of the answers are offered in the papers on agent causation that are cited above, although the sufficiency of the answers is disputed. But subject origination can remain untouched by the objections, because there is no claim that the concept picks out something in the world over and above the objects, the properties and powers of objects and the events involving objects that are picked out by the natural sciences. Use of the concept is merely an alternative way of looking at the reality to which the natural sciences apply the concept of event causation.

This is also the reason why the use of two different conceptions of ourselves, the naturalistic conception and the originator conception, does not give rise to the problems that can be associated with having two different explanations of the same event. We may explain an event in terms of the world's going on in accordance with the laws of nature, and we may also explain the event in terms of the subject's having deliberated in a way that is implicitly seen as having involved subject origination. The problems that can arise from having two different explanations are discussed by Jaegwon Kim in "Mechanism, Purpose, and Explanatory Exclusion". He sets out a metaphysical principle of explanatory exclusion, that two explanations "can both be correct explanations only if either at least one

of the two is incomplete or one is dependent on the other" (ibid., page 275). He adds an epistemological principle, that "no-one may accept both explanations unless one has an appropriate account of how they are related to each other" (ibid., page 275). The user of the two conceptions of ourselves can comply with the epistemological principle. The account of the relationship between the two explanations of a given instance of deliberation is located not at the scene of the deliberation, but way back in the idea of selecting conceptions to use in order to make sense of the world and of our lives. Furthermore, the user of the two conceptions need not worry about falling foul of the metaphysical principle. When the concept of subject origination is used, that is not in itself an attempt to explain why one action was performed rather than another. Contrastive explanations are not provided, and to that extent explanations in terms of subject origination are incomplete. The user of the two conceptions remains subject to any difficulties that may be associated with offering an explanation of the outcome of a deliberation in terms of the subject's reasons for selecting options, as well as an explanation in neurological terms. But he is not subject to such difficulties on account of his use of the two conceptions, except in the indirect sense that a user of the originator conception in the context of deliberation sees the subject as weighing and using evidence, so that he can hardly avoid giving explanations in terms of reasons.

One specific objection to agent causation does, however, merit particular attention, because it appears to present a strong challenge to the idea that the concept of subject origination can do the work that it is intended to do. This is the objection that the agent might lack adequate control. As noted in section 2.3, I can only issue a promissory note that is not backed by any gold. I simply assert that we should see ourselves as deliberating with mastery, free but in control. But if we merely see ourselves as engaging in subject origination, it is not self-evidently unacceptable to see ourselves as in control. Timothy O'Connor writes, "Agent-causal events are intrinsically actions – the exercising of control over one's behaviour. It is senseless to demand some further means of controlling this exercise of control" (*Persons and Causes*, pages 58-59). This is a contentious claim if one wishes to integrate the exercise of control into a naturalistic conception

of the world. It is much more secure if one has no such ambition.

I shall return to the relationship between my approach and agent causation in section 3.8, once I have set out a concept of the subject that accommodates subject origination. The fact that the concept is not one of a subject that is a worldly addition or enhancement to the physical person has a bearing on the relationship.

The intentional stance

I shall now discuss the intentional stance as presented by Daniel Dennett, the stance from which people are regarded as rational agents with beliefs and desires (Dennett, *The Intentional Stance*, pages 15-35). I shall read Dennett as advocating a type of intentional stance that could be set out using only the resources of the naturalistic conception of ourselves. In particular, our sense of what it was like to have certain experiences would, on my reading, have no essential role to play. References to Dennett should be read as references to Dennett as interpreted in this way. I cannot be sure that the real Dennett would agree with this reading, because he does not tackle the issue in precisely these terms. But passages in which Dennett discusses issues in the area of simulation give no reason to think that my reading of him is mistaken (*The Intentional Stance*, pages 53-54 and 98-101). If we adopt such a naturalistic reading of the intentional stance, that marks a difference from the originator conception. I shall ask whether, despite that difference, we would still need to rely on inner experience in order to identify the intentional stance that was appropriate to a given set of creatures.

I shall also refer to *the* intentional stance, meaning the stance in general, and to *an* intentional stance, meaning a particular way of interpreting behaviour by using a particular set of concepts of beliefs, desires and so on. I shall take it that an intentional stance is a correct one for a set of creatures, although not necessarily the only correct one, when it allows its users to systematize and predict the creatures' behaviour by attributing thoughts that explain the occurrence of certain behaviours. Stances may therefore possess degrees of correctness. The requirement to allow systematization and prediction means that an intentional stance must allow both the thoughts that are attributed to creatures, and their

behaviours, to be described using a modest number of central concepts which fit together in a structured way. A stance that merely allowed the re-statement in an intentional vocabulary of all of the micro-physical detail of the world would not be any good.

I shall discuss the community's route to an intentional stance. I shall leave to one side the development of the individual, and how the individual comes to grasp an intentional stance that is used within the community. This focus on shared intentional stances tends to induce a theory theory approach to folk psychology, rather than a simulationist approach. If we reflect on the shared intentional stance of a community, we are very likely to state it in theoretical terms. But I do not argue for theory theory at the level of the individual. That which may best be captured in theoretical terms when describing a community, could still be implemented in a simulationist way in the individual members of that community.

Use of the originator conception is more daring than the adoption of an intentional stance. An intentional stance could perfectly well be applied even if one saw humanity in purely naturalistic terms. Using the naturalistic conception, it is perfectly possible to say that someone works extra hours because he is ambitious and wants to get promoted quickly, just as it is possible to say that the climate in north-western Europe is heavily influenced by the Gulf Stream. Both remarks involve a high degree of abstraction, but there is no need for naturalistic descriptions to be given in terms that reduce everything to individual particles. A naturalistic reading of the first remark would not, however, capture any sense of what it was like to choose actions out of ambition, to give up hobbies and to put oneself under pressure in order to have a successful career. In order to have a sense of that, one would have to be a human being who could experience such feelings as ambition. That would go beyond the intentional stance as presented by Dennett.

This difference between Dennett's intentional stance and my originator conception neutralizes what would otherwise be a source of disagreement between us. In chapter 7 of *The Intentional Stance*, Dennett finds a place for the intentional stance in the conduct of science and in the expression of scientific results. I agree that it has such a place, even though neither our sense of being points of origin, nor our sense of what it is like

to have given motivations or emotions, can feature in any statement of scientific results. Such senses are tinged with the specifically human experience of life, so they could not be grasped by a much wider range of rational beings. It is, therefore, essential to the relationship between Dennett's intentional stance and the sciences, that our senses of being points of origin and of what it is like to have given motivations and emotions fall outside the scope of the intentional stance.

While I am in extensive agreement with Dennett, I do not regard his intentional stance as free-standing. Our adoption of it depends in practice on our inner experience, in a stronger sense than he allows (*The Intentional Stance*, pages 53-54). The reason is that before we can take an intentional stance, we must identify the specific types of belief, desire and the like that it is appropriate to attribute to the creatures concerned. We use our inner experience to identify the concepts that we should use. If we did not have that inner experience to guide us, we would be faced with too wide a range of possible sets of concepts to use. Furthermore, it needs to be our inner experience that we use, our experience of having experiences, of deliberating, of deciding and of acting, rather than our experience in general. We can only identify the concepts to use if we stand back from our ground-level experiences, such as the experience of seeing a tiger, and consider what it is like to see a tiger, and the associated feelings of fascination and of fear. My concern is that while the patterns of conduct that are revealed by taking an appropriate intentional stance may or may not be real, it would not be at all easy for a detached observer, without the benefit of inner experience in common with the creatures observed, to identify any such appropriate stance. (Different, but related, concerns are set out in Haugeland, "Pattern and Being".)

There is a way to avoid making the identification of an appropriate intentional stance dependent on our inner experience. We could start with outward behaviour and changes in internal physical states, and guess which concepts to use in order to take an intentional stance. But we would be very lucky to find the right set of concepts in that way. It is not easy to imagine how difficult it would be to find the right set of concepts, because our knowledge of our own needs, desires, and ways of thinking and acting is deeply ingrained in us. If we imagine being confronted with aliens, and

trying to take up an intentional stance toward them, that can give some impression of the difficulty. The logic of this way of looking at the problem is that we could not draw on our own experience, because we would have no reason to expect the aliens to have inner experience that was anything like our own. We would be thrown back on outward behaviour and on changes in internal physical states as our only resources. If, for example, the aliens often came close to one another, we could not assume that this reflected a need to bond with one another or to reinforce hierarchies. We could not even assume that they had a propensity to promote their individual survival, although it is unlikely that they would last long enough for us to meet them if they had no propensity to behave in ways that at least did not impede the survival of their species. If we found a set of concepts that did allow us successfully to systematize and to predict their behaviour, that would be evidence that we had got it right, but we could expect to spend a long time guessing before we had such success. This is enough to make the dependence of the identification of an appropriate intentional stance on our inner experience a fact for all practical purposes. There is, however, one circumstance in which this conclusion would not hold. If the aliens had outward behaviour that was very similar to our own, we could use the intentional stance that we used toward human beings. We would then succeed in systematizing and predicting the aliens' behaviour. Correspondingly, the argument about aliens does not show that our inner experience has any essential role in extending our existing intentional stance to new human beings whom we meet, because their outward behaviour is usually like our own. But as will be argued below, this does not exclude a vital role for inner experience in our first formulating an intentional stance toward human beings.

Another way of making the point is to consider the problem of radical translation, the problem of making sense of a language when we have no starting point and have to work entirely from linguistic and non-linguistic behaviour, and the related problem of radical interpretation, the problem of trying to make sense of the speakers of an unknown language by attributing mental states to them. Quine and Davidson argued for the importance of the principle of charity, a principle that has various formulations but the core of which is an assumption that speakers of a

language mostly tell the truth and mostly get things right. In practice, this assumption means that we would try to interpret an unknown language in ways that would maximize agreement between its speakers and ourselves. A useful supplement is a principle of humanity. This principle invites us to assume, so far as we can, that the speakers of an alien language are like us. This supplement is arguably essential when the problem is that of radical interpretation. (A brief account of the early history of both principles, with references, is given in Dennett, *The Intentional Stance*, pages 342-344.)

We could not legitimately apply either principle to any significant extent when trying to identify an appropriate intentional stance toward aliens. Starting with the principle of charity, what counted to aliens as getting things right might be very different from what we would regard as getting things right. What was salient for them might be very different from what was salient for us. Convincing examples cannot be offered because they would have to be incomprehensible to us, but we can get an idea of one direction in which such examples might lie by considering Wittgenstein's articulate but incomprehensible lion (*Philosophical Investigations*, part 2, section 11, page 190). For us, the salient facts about some animals in the bush are the number and species of the animals. For a lion, the salient fact might be the ratio of the number of lions that could be fed by killing the animals to the effort that would be required in order to kill them. This is not a wholly convincing example, because we could calculate the ratio in which the lion might be interested by finding out how many animals, of which species, were present. But if I could give incomprehensible examples, that inferential link would be broken. Turning to the principle of humanity, any assumption that aliens were like us would be wholly unfounded.

It might appear that we could at least count the conceptualization and recording of the environment in a way that would promote the survival of the species as getting things right, but even if we could make that assumption, it might provide very little help in the interpretation of aliens' behaviour. It would be especially unlikely to be helpful if they lived in an environment where survival was not difficult because there were hardly any threats, so that an imperative to promote the survival of the species could not be seen as a significant motive for much of their behaviour. I therefore contend that there are narrow limits to how much we can be confident

would be achievable by considering what beliefs and desires a being ought to have, given the shape of its ecological niche (compare Dennett, *The Intentional Stance*, pages 49-50).

We should also consider the global version of charity for which Karsten Stueber argues (*Rediscovering Empathy*, section 2.1). In his words, "Radical interpretation implies only that to justifiably attribute a belief, an interpreter must be able to attribute a sufficiently large enough belief system with a certain amount of structure and complexity" (ibid., page 76). While this version would appear to be immune to the objection to the use of the principle of charity that I have raised, it would unfortunately not help us in interpreting aliens. We would not have enough clues to allow us to attribute any large system of beliefs that was usefully structured. We might manage to attribute some large and complex system on the strength of patterns of cerebral and external activity. But we would probably only be able to attribute a system that was not well-structured using a modest number of intentional concepts, because we would not be able to identify the key concepts. That would make the system useless in the systematization of behaviour. It would also be useless in the prediction of behaviour, because the data that were available to a predictor would be limited and would be given in terms that used broad-brush behavioural concepts. Those data could not then be related systematically to the attributed complex system of beliefs. For me, such difficulties matter. For Stueber and others, including Davidson, the point of reflection on radical interpretation is not to interpret aliens, but to use reflection on that extreme task in order to understand what is involved in the interpretation of human beings. For me, what matters is the contrast between the interpretation of aliens and the interpretation of human beings. We need to put the two tasks side by side, giving them equal prominence, in order to see how important it is that the interpreter be of the same kind as those whom he interprets, in our case the human kind.

There is an apparent refutation of the conclusion that the identification of an intentional stance depends on access to inner experience of the appropriate type. This is that we do appear to succeed in making sense of the behaviour of animals, when we use some of the resources of our intentional stance toward human beings. We can see

animals, and especially our domestic pets, as exhibiting such things as joy, determination and defensiveness. We can use such concepts to interpret their behaviour, and to relate it to goals that we see them as having. We must be wary of misinterpretation, driven by an inclination to anthropomorphize, but we do seem to get somewhere, even though animals are very different from human beings and probably have very different inner experience, to the extent that they have inner experience at all. If we can get that far with animals, why should we not be able to get somewhere with aliens, merely on the basis of their outward behaviour and changes in their internal physical states, without having to spend a long time guessing the appropriate concepts to use in formulating an intentional stance? My response is that while animals differ from us, they do have quite a lot in common with us, including needs to eat, sleep, maintain a comfortable temperature and reproduce. We have also fitted some of them, especially our domestic pets, into our own lives, giving several points of contact between our way of life and their ways of life, and we have chosen and bred species that can fit in well with our lives. Our pets may be mentally distant from us, but they are not so distant in their form of life as we should expect aliens to be, save that our pets are very distant from us in their level of mental sophistication, while aliens might be closer to us in that respect.

Another argument can be brought to light by considering how we might have arrived at a satisfactory intentional stance toward human beings. Our inner experience in itself may well have no intrinsic, necessary structure. Such a lack might be thought to imply that we would have had to develop a theory of our own inner experience, structuring it in some way, before we could have used that experience to help us to identify the correct concepts to use in formulating an intentional stance toward human beings. But it seems more likely that we would have had to structure our inner experience and people's behaviour in tandem, because reflection on the set of inner experiences would not by itself have yielded principles that would have determined an appropriate structure for that set. Inner experience might or might not then have had an essential role to play in determining the appropriate concepts to use in categorizing and structuring behaviour. That is, it might be that we could have structured the behaviour without reference to the inner experience, although we might still have

needed reference to changes in internal physical states. If inner experience were dispensable in that way, then the formulation of an intentional stance might not depend on inner experience after all. An appropriate structure of behaviour, given in terms such as those of meeting needs, avoiding dangers and complying with social norms, would directly yield an appropriate structure of beliefs, desires and intentions. That would suggest that even when faced with aliens, we might be able to manage without access to their inner experience.

We should start with the question of why behaviour is amenable to any straightforward systematization at all. The answer is likely to be that the Universe is a sufficiently inhospitable place that only creatures with elaborate patterns of behaviour that embodied consistent responses to their environments would be able to evolve to a high degree of sophistication. This thought directs our attention to a specific way in which we might hope to manage without access to inner experience, and work merely from behaviour and from changes in internal physical states to an intentional stance, without spending ages guessing which intentional stance to apply to some aliens until we found that one of our guesses worked. The clue is in the phrase "consistent responses". In order for us to see responses as consistent, we would need to have an appropriate way of classifying states of the environment, so that we could say which states were sufficiently similar to make sameness of response explicable. Such a classification would be based on the needs of the creatures concerned, and on the perceptual apparatus that they used. Thus if a creature was of a kind that needed water regularly, any state of the environment that involved a shortage of water, however caused, should lead to water-seeking. Similarly, if a creature was of human size and physical frailty, and also of a kind that gathered information about immediate dangers visually, then any state of the environment that produced the same condition of the visual cortex as a charging bull, including a hallucination, should lead to a sharp movement sideways when the real or apparent bull appeared to be too close to be able to change course. If we were to classify states of the environment in terms that reflected both the needs of the creatures and their perceptual abilities, we might be able to identify appropriate sets of beliefs, desires and so on by analysing their physical needs and their perceptual apparatus.

We cannot, however, be sure that this approach would work. First, as already noted, aliens might have needs that were so different from ours, and that were defined by reference to such different ends, that we would not get far in identifying those needs. Second, while we might be able to analyse an alien's perceptual system from the outside, and work out that it had, for example, a certain power of discrimination between different patterns of electromagnetic radiation, or between different patterns of physical pressure, we would not know what the perceptions signified to the alien. Sometimes, there would be behaviour that we could correlate with the perceptions in order to get some idea of what was signified. If, for example, someone sits calmly until he is given an unusual stimulus, and then immediately runs from the room, it is reasonable to conclude that he takes the stimulus to indicate the presence of something that he wishes to avoid. But sometimes, there would not be behaviour that would allow us to deduce the significance of perceptions. A perception might have a significance which meant that it gave the perceiver new dispositions that were never actualized, because the appropriate conditions never arose. Thus someone might see a high mountain, decide that it looked like a good place to go when seeking calm and detachment from everyday life, but then never feel the need for a period of calm and detachment. In other cases, there would be behaviour that we could only use to deduce the significance of perceptions if we already had a full understanding of the creature's way of life, an understanding that we would not have if we were still seeking an appropriate intentional stance. There are perceptions that have delayed effects on behaviour, often in connection with other perceptions and in ways that we only notice because we do have a full understanding of our own way of life. A woman may see a display of watches in a shop window, then two days later notice some birthday cards in another shop, remember that her son has a birthday soon and go back to the first shop to buy him a watch. An alien could only link the perceptions to the behaviour, and make sense of the whole pattern, if he already knew about the significance to us of family relationships, of birthdays and of present-giving. We would likewise be unable to work out the significance to aliens of their perceptions, merely on the basis of their behaviour and of changes in their internal physical states, except in the simplest cases such as that of

suddenly running from the room when given a particular stimulus, unless we already understood their way of life. It may well be true that members of any species that has evolved to a sophisticated level have consistent responses to states of the environment, with those states of the environment being classified in a sophisticated way by reference to the creatures' needs and their perceptual abilities. But it does not follow that we could reliably work out what was significant to them, and determine an appropriate intentional stance, on the basis of our observations of their apparent needs, their perceptual apparatus and their behaviour. Without an understanding of what was significant to them, we would spend a long time guessing until we found an intentional stance that was correct, in the sense that it allowed both the systematization and the prediction of behaviour.

Turning to ourselves, the task is much easier because we do have access to our own inner experience, and because we are well aware of our own needs and of our way of life. Thus we have a direct route to an understanding of the significance of different perceptions, and a basis on which we can construct an appropriate intentional stance toward human beings. We are directly aware of the devices and desires of our own hearts, and that is why we can interpret other people. But even with this advantage, it is still not clear how we could first have arrived at the rich intentional stance that we now use so effortlessly. After all, our inner experience had to be structured, as well as our outward behaviour.

A key point is one that applies equally to biological evolution. There is no need to leap straight from the simplest to the most complex, whether from single-cell life forms to human beings or from practically no intentional stance to a full stance. Evolution depends not just on chance, but on chance plus selection. Small changes that confer some advantage can become securely embedded, and can provide jumping-off points for the next round of small advances (Dawkins, *Climbing Mount Improbable*, chapter 3). We see a parallel effect in theoretical studies. There was, for example, no single leap from the mathematics of Pythagoras to the mathematics that we have today, but a long sequence of new ideas, each thoroughly tested as it was put forward and then incorporated into the mathematical tradition if it was recognized as an advance, or discarded if

it was not. There is no reason why the development of our intentional stance toward human beings should not have benefited from such an effect. There is even specific support for such an effect in the interdependence of theory and inner experience. As we evolve an intentional stance, or any other folk-psychological understanding of human beings, it starts to feed back and shape our inner experience. Once theory develops to a new level in a way that roughly fits inner experience that has not already been theorized to that level, we pick out the inner experience in a way that fits the new bit of theory. That helps to give the new bit of theory a firm foundation. We can, for example, have the experience of postponing immediate gratification in order to gain some advantage later on. We can be aware of it as that experience, with its own distinctive phenomenology, once we have a corresponding level of psychological understanding of what human beings sometimes do. We can also identify a mechanism of transmission of an evolving intentional stance in the stories that we tell, and a mechanism of development in that those stories can be changed and can become more sophisticated. That development is greatly accelerated when people become conscious of the role of stories in capturing psychological insights, and start to write stories in order to make particular psychological points. The role of stories is set out by Daniel Hutto in *Folk Psychological Narratives*. There is also adaptive advantage in developing a sophisticated intentional stance. The most unpredictable things in the world are our fellow human beings. If we can understand them, we can both work with them and work around them. That greatly enhances our ability both to survive and to prosper. Finally, a gradual evolution of an intentional stance could have had a starting point in the most basic experiences and drives, as when pangs of hunger lead us to seek food.

I stated above that we might have to determine the structures of our inner experience and of our behaviour in tandem. That thought raised the question of whether we might be able to manage without reference to the inner experience of aliens when working out how to structure their behaviour. We probably would need to structure our own inner experience and our own behaviour in tandem, because the range of different structures into which experiences on their own might fall would probably be too wide. But this does not mean that inner experience would be redundant. The

development of the two structures in tandem would be guided by the need to arrive at a reflective equilibrium between them. A necessary condition for equilibrium would be that inner experiences, and the behaviours that one would correlate with them on the basis of parallels between the two structures, should be apt to one another. They would need to be apt over extended sequences, and not merely at the level of momentary experiences and behaviours. Thus, an inner experience of slight anxiety on account of a distant threat should be correlated with the behaviour of keeping an eye on the threat, an inner experience of greater anxiety on account of a reasonably near threat should be correlated with making preparations to avoid the danger, and an inner experience of fear on account of an immediate danger should be correlated with immediate evasive action. That is, we should expect a rising scale of concern to parallel a rising scale of preparedness. The concept of making preparations, with its associations of planning, of timescales, of foreseen needs and of motivation by awareness that the unprepared may suffer, could then have a place in the intentional stance that we would adopt toward human beings. Such an exercise in establishing the aptness of correlations that were given by parallels between the two structures would take us beyond the correlations between momentary experiences and behaviours that could be apprehended directly, without formulating structures, such as the correlation between fear and running away. There would indeed be a three-way reflective equilibrium between the structure of inner experience, the structure of behaviour and the content of the intentional stance, with the content of the stance embodying the aptness of the correlations that were given by structural parallels. While we can legitimately see ourselves as developing an intentional stance by reference to inner experience and behaviour, the fact that unique structures of inner experience and of behaviour, and therefore unique parallels, would not be forced on us by the natural facts, means that we should not see a one-way relationship of dependence, in which an intentional stance emerges from, but has nothing to contribute to, the structuring of inner experience and of behaviour.

I do not claim that the equilibrium constraint would uniquely determine either the structure of experience or the structure of behaviour. But I do claim that inner experience would have an important role to play

in identifying what it was about behaviour that made it apt to external circumstances, and hence in influencing the choice of concepts to use in the intentional stance, concepts that would make clear how we should systematize whatever behaviour we observed. The inner nature of the experience, its presence in the inner life of the subject, is vital here. A perception that is described in terms of what is being perceived, or in terms of the states that it produces in perceptual areas of the brain, is not of obvious significance. Why should an approaching tiger matter? The significance becomes obvious if we ask the subject how he feels about the tiger, but only if we can grasp the point of what he says. We can do that because he is another human being, and is therefore sufficiently like us. The significance might also become obvious if we noted that areas of the brain that corresponded to certain preferences or emotions were activated, but only if we knew that activity in those areas of the brain corresponded to those preferences or emotions. We would only know that if experimental subjects had reported their thoughts and feelings at times when we knew that those areas were active, and if they had given those reports in terms that we, as human beings, could understand. Thus there is no getting away from accounts that are given in terms that reflect the specifically human experience of life. I have not shown that it would be impossible to arrive at a structure of behaviour that was adequate to yield an appropriate intentional stance without reference to the inner experience of the creatures concerned, but I hope to have shown that reference to inner experience is the high road to an intentional stance, and that any other road would be a long and winding one.

We should also consider specifically simulationist approaches to folk psychology. Several approaches come under this heading, but they all involve ways of understanding human beings that are only likely to be available to those who have been brought up in human communities, and who have human inner lives. (Correspondingly, a simulationist approach might be used by an alien, but only to understand aliens of the same type.) Jane Heal sets out the need for a simulator's cognitive apparatus to be sufficiently similar to that of the person simulated ("Understanding Other Minds from the Inside", page 30). Alvin Goldman acknowledges that the use of simulation to predict what someone else will do capitalizes on

similarities between the minds of the simulator and of the person whose actions are to be predicted (*Simulating Minds*, page 20). Karsten Stueber's *Rediscovering Empathy* gives a central role to re-enactive empathy, something one would only be likely to have if one shared enough experiences with the person who was being understood. Daniel Hutto's *Folk Psychological Narratives* sets out a process of listening to stories as a way of acquiring folk-psychological competence. While an adult outsider could read the stories, he would need the right concepts in order to distinguish their significant content from insignificant details, and in order to understand the significant content. So this story-telling account would not diminish the importance of having appropriate inner experience. The points that are made in this paragraph about the need for mental similarities between simulator and simulated are, however, only points about what makes simulation feasible. So they do not contradict the point made in section 3.3, that simulation is not essentially a matter of having empathy. Nor indeed would mental similarity, or reliance on it when simulating another person's mental processes, entail the presence of empathy in the sense of sharing a feeling.

The practical need for those who would take up an intentional stance to make use of the resources of their inner experience, and for that experience to be at least roughly like that of the objects of the stance, extends to functionalism. Functionalists need to identify the functions that are to be performed by different states of the brain. The state of fear, for example, has the function of leading to heightened awareness, and to increased readiness to flee. We ask what functions need to be performed, in order for a being to lead a life of a given type. If that type is defined in intentional terms, we will in practice need to make use of our own inner experience in order to define the type, so that we can identify the functions. (This conclusion does not follow if we define the type of life in non-intentional terms. We may, for example, define the type of "life" of a robot non-intentionally. But we then individuate the robot's behaviours, and identify the functions to mention in a functionalist analysis, by reference to the list of tasks for the robot to perform that has determined the robot's design.)

Giving a more substantial role to our inner experience might address some of the difficulties that functionalism faces. We could allow ourselves

to see the functions as essentially infused with the phenomenal nature of the experiences that allowed the identification of those functions. The difficulties in question are those that are captured in the thought experiments of the Chinese room, the China brain and the inverted spectrum. One reason why they count as difficulties is that we want a philosophy of mind to be adequate to our experience of life. Functions look as though they can be defined in purely structural terms, linking perceptions and behaviour, without any specific experiential content, whereas our lives are full of experiential content.

Such a move would be radical. It would give inner experience a role not merely in identifying an appropriate functional analysis, but in formulating its content once it had been identified. That would in turn threaten the notion that there was a place for a functional analysis in the expression of scientific results, because those results need to make sense even to beings who do not have inner experience that is anything like our own. The corresponding place of the intentional stance in the expression of scientific results would likewise be threatened by giving such a role to inner experience. I do not advocate this radical move, but I shall now use the thought to pose the question of whether a detached psychological theory, one that held well back from any such move, would be adequate.

Detached psychological theories

A detached psychological theory is one that is formulated in terms that are independent of the phenomenal nature of our experience, not just our sensory experience, but also our experience of thinking and of acting. It is detached in the sense that an observer who was not a participant in human life, an alien who was psychologically detached from us, could grasp it. What would we think of such a theory?

The argument so far has only shown that the resources of our inner experience would in practice be needed in order to identify the appropriate functions or psychological concepts, in terms of which to formulate a theory. Once they had been identified and assembled into a theory that allowed us to systematize and predict people's behaviour, the methods that had been used to identify them would become irrelevant to their appropriateness. The theory might well be capable of application, in a

scientific way, by a detached observer who sought to systematize or to predict human behaviour. The observer would not need to rely on the deliverances of his own inner experience, or on his inner experience's bearing any relationship to the inner experience of human beings. The only requirement would be that the observer had access to the necessary data. Those data might include states of the brains of subjects, and what subjects would say in reports of their internal reflections if they were asked, as well as the outward behaviour of subjects.

The significance of the practical need for reference to our inner experience when constructing a psychological theory can indeed be limited in this way. A detached observer might work like that, although it would not all be plain sailing. Much would depend on the type of psychological theory that was involved. A humanistic psychology, for example, could not be used in such a detached way. Furthermore, the centrality of the fact that our thoughts are about things might mean that there was an obstacle to the existence of a wholly successful detached observer. This would be the obstacle that is identified by Galen Strawson's argument that intentionality that is attributable merely on the basis of behaviour, without reference to experience, cannot be full intentionality (*Mental Reality*, section 7.8; the point is also argued in relation to the intentionality of emotions in Goldie, "Emotions, feelings and intentionality"). But even leaving aside such restrictions and obstacles, if we pursue the idea of a detached observer who uses a theory that is decoupled from inner experience, we find that we are led back to the significance of inner experience by a different route.

It is tempting to think that the methods that were used to identify the concepts to use in a theory would automatically be irrelevant, once the concepts had been identified. An obvious analogy is that while Newton may have been inspired by the fall of an apple, his theory of gravity would have been exactly the same even if no apples had ever existed after that moment of inspiration. The analogy is, however, misleading. The content of a theory of gravity is unaffected by what types of organic matter, if any, happen to exist. The significance of a theory of gravity is likewise unaffected, except in the indirect sense that gravity helps to determine which types of living thing can exist and what they can do, and that it will only have that influence if living things do exist. The content of the

psychological theory that we may envisage being used by a detached observer might likewise be unaffected by the nature of the inner experience of the observer, or of the human beings to which it applied. But its significance certainly would be affected by its detachment from the inner experience of human beings.

The theory that is envisaged would simply be a scientific theory that allowed its users to systematize and to predict human behaviour. It would set out how certain states and behaviour were typically followed by other states and behaviour. It might also set out deeper structures that explained such surface regularities. There would be no need for the theory to claim perfect regularities. It might say that a given state, in given circumstances, was followed by given behaviour 70 per cent of the time and by different behaviour 30 per cent of the time. That is, seeing our choices and behaviour in the terms of such a theory would not necessarily take away our sense that we could do things other than the things that we actually do. But there would be a clash with our sense of how we deliberated. The theory would identify all of the sources of the outcomes of our deliberations. The sources would include our desires and our thoughts about evidence, seen as operating within a context of character traits rather then as physical causes. That is, a detached psychological theory could speak of occurrent thoughts and standing dispositions as causes, without reference to the electro-chemical events and states that were their physical counterparts. (The contents of the thoughts and dispositions would have to be given in terms that would strike us as inadequate, because the terms would be divorced from our inner experience, but the contents could still be given in terms that were sufficient to allow the systematization and prediction of behaviour.) These sources of the outcomes of our deliberations would be the things from which the outcomes followed, whether with perfect regularity or merely with statistical regularity. The theory would therefore omit any notion of deliberation as more than following a program, because it would find no place for the stage of standing back and reviewing options at the selection stage in a way that was independent of antecedents, and that would allow for a controlled and explicable break with those antecedents. Such a theory would to that extent not be adequate to our inner experience, even though a detached observer who did not have inner

experience that was like our own might well regard the theory as adequate because he would have no grasp of what was missing. Furthermore, the phenomenological inadequacy would exist whether or not one considered that the adoption of a view of deliberation as involving mastery and subject origination was the appropriate way to accommodate that phenomenology.

This argument applies to psychological theories, regardless of how they are created. It would therefore apply whether the observer had identified the concepts to use through a grasp of human inner experience or through lucky guesswork. The argument is tinged with circularity. A detached psychological theory would not strike us as adequate, but our sense of its inadequacy would depend on possession of our type of inner experience. The shape of the argument is, however, more that of the first turn of a spiral than that of a flat circle. We start from the phenomenology. The felt inadequacy of a detached psychological theory that shows no trace of the phenomenology drives us round to take the phenomenology seriously, as something that should be reflected in our psychological theories. So we do not end up back where we started, at the fact of the phenomenology and a mere feeling toward it, but one level above that starting point, with an articulated desire to take the phenomenology seriously.

This conclusion does not undermine the view that there are places for the intentional stance and for functional analyses in the conduct of science and in the expression of scientific results. We can expect to need to go beyond the scope of the sciences in order to identify an appropriate intentional stance or functional analysis, but once we have done so we can use that stance or analysis within the scientific enterprise. Nonetheless, a psychological theory that was limited to the use of an intentional stance or of a functional analysis in that scientific way would strike us as inadequate, because it would be a detached theory (compare Goldie, "Emotions, feelings and intentionality", pages 248-250).

3.8 A concept of the subject

We need a concept of the subject that will allow us to see ourselves as

engaging in subject origination. Such a concept of the subject is, like the concept of subject origination itself, initially pressed on us within the narrow field of choices of action and adoptions of belief where systematic reflection is to be expected. But if we secure the concept within that narrow field, we can then use it more widely.

Setting out the concept amounts to elaborating the originator conception. A concept of the human being as a mere object that was within the world, even a very sophisticated object, would not suffice because a subject that was seen in that way would be seen simply as falling within the world's causal network. The subject could be seen as a locus of events, but not as an originator of them, because there would be no way to see it as generating causally efficacious extraneous interventions.

Another constraint is that our concept of the subject must locate each of us in space-time. Location is fundamental to our lives. We can only experience things to the extent that information about them reaches our locations, and we can only do things to the extent that we can influence the world from our locations.

Finally, we want a concept of the subject that will minimize any mystery. Mysticism may be a wonderful source of life-changing experiences, but mystery is an unwelcome element in sober philosophical theories. The concept that is put forward here unfortunately does involve a little bit of mystery, because of the need to have a concept of the subject that does not lead to our seeing it merely as an object that is within the world. Fortunately, the mystery does not intrude on the province of the natural sciences.

I shall use the term "boundary concept" for the concept of the subject that I offer. Its inspiration is a remark that was made by Wittgenstein, "Das Subjekt gehört nicht zur Welt, sondern es ist eine Grenze der Welt", "The subject does not belong to the world. Rather, it is a boundary of the world" (*Tractatus*, 5.632). I take this remark as my starting point, and do not commit myself to the use that Wittgenstein makes of it, nor to the uses that other philosophers have made of it. I am not, for example, concerned with the question of whether "I" is a referring expression. But I shall take Wittgenstein's words at face value. He says that the subject is a boundary of the world, not that it is located at a boundary.

I shall now elaborate on the concept of the subject as a boundary of the world, the boundary concept, in two ways. I shall first develop a visual image that may be useful, although we must be wary of drawing conclusions on the basis of any such image. Then I shall develop the logic of the concept.

The image is that of the boundary of a small hole in the world. I shall take the hole to be a three-dimensional region of space, so that the boundary is a two-dimensional surface that completely encloses the region, and that also persists through time. The surface is located where a person's body is located. It is an interior boundary of the world, not a boundary that runs right round the outside of the world. That which is below the surface is off limits, not part of the world. The surface itself is not within the world, because something is only within the world if there are points that are parts of the world on both sides of it. The image must not be put under too much pressure. If the interior, the hole, were not part of the world, one could cause difficulties by asking whether either that interior or its boundary could even be spatial. I shall not, however, take up such concerns. The image is only meant to be rough and ready, and I shall not rely on it in the argument. It also does not matter to the argument whether we take the world to include or to exclude the points that constitute the surface itself, although I think that the image is more satisfactory if those points are taken to be included, as the last points that are parts of the world.

Three choices of surface naturally suggest themselves. The first is the skin. The second is the surface of the brain. The third is a tiny sphere. None of these choices is ideal, but each is suited to certain circumstances. Most of the time, in our ordinary encounters, it is natural to feel that the limit of a person is his or her skin. But doctors regard what is under the skin as within the world, to be studied and acted on like any other part of the world. For them, any boundary of the world must retreat so as to leave the body within the world. It would be natural to take the subject back to the surface of the brain, given that thinking, feeling and deciding are concentrated in the brain. But even that would not always be appropriate. There are doctors and others for whom the brain is very much within the world, with all of its parts amenable to study and to treatment. One could question whether the concept of the subject as an entity that was not to be

analysed had any role to play when the brain was analysed in such a way, given that our sense of self depends on several parts of the brain working together, and that there is no one kernel of grey cells that can be said to be the self. But to the extent that there was a role for the concept, one could take the subject to be the surface of a tiny sphere, with all of the physical parts of the body outside it. Alternatively one could abandon the pursuit of an appropriate image at this point, on the basis that it was not sensible to identify the subject with the surface of a region that was outside the world, while simultaneously seeing the subject as a product of the interaction of sub-personal systems that were within the world. There is a specific reason why it would not be sensible to continue to pursue an appropriate image. That which is below the boundary is to be treated as off limits. If the boundary can be identified with the surface of the body or of the brain, then all or most of what is in fact the causal mechanism of thought is screened off. That makes it easy to overlook the causal processes that take place. It is then easy to overlook the causal closure of the physical, because gaps can be seen in the physical causal network. That in turn allows us to see deliberation as conducted with mastery, and as involving subject origination. (This is only meant as an indication of how the image could be taken. It cannot be used in argument, because to do that, one would also have to establish that all and only the right parts of the causal mechanism were screened off.) If, on the other hand, the image leads us to see nothing physical as screened off, because the boundary is reduced to the surface of a tiny sphere, then this connection between the content of the image and the idea of overlooking certain causal processes is lost.

Fortunately, the image of the subject as a surface that is a boundary of the world is no more than an image. We are free to vary the image, choosing whatever surface may be appropriate to the occasion, or not to bother with the image. The image is put forward to capture the idea that the subject is not wholly embedded in the causal network, because there are no parts of the world on the far side of the surface, but equally that the subject is not part of some mysterious other world. I shall now analyse this idea, leaving the image behind.

The great master of things that are not fully represented by the phenomena, and that on some but not all interpretations form a separate

world from the phenomenal world, was Immanuel Kant. The noumena, things as they were in themselves, were placed off limits. The concept of a noumenon was a mere limiting concept, a way of identifying the unknowable without saying anything about it. By trimming the ambitions of a realist view of the world, Kant could avoid both a paralysing scepticism and the otherwise devastating conflicts of transcendental ideas, the antinomies. The subject, the "I" in "I think", was also kept apart from the phenomena, for different reasons. Doing so made possible the freedom that was essential to Kant's view of human beings as potentially moral creatures. This concept of the subject also had to be a mere limiting concept. The subject had to be something that was inferred from its role and that was not visible in itself, both for the sake of freedom and in order to avoid entanglement in the mistaken inferences of rational psychology. Only the minimal properties that followed from the subject's role could be attributed to it. (The key passages in the *Critique of Pure Reason* are A254-255/B310-311 on the concept of a noumenon, A345-348 and B403-409 on the subject as thinker and no more, and A532-558/B560-586 on freedom.) The final product was by no means free of tension. As Jerrold Seigel puts it, "Kant both sought and avoided the unification of the empirical and intelligible forms of the self" (*The Idea of the Self*, page 316).

The subject as it is conceived using the boundary concept can be taken to be a replacement for Kant's intelligible subject. It is likewise a being that is not to be understood in terms of physical mechanisms, but with the difference from Kant's intelligible subject that it has a location in an independent and pre-existing spatio-temporal world. I shall return to the idea of locating the subject in that pre-existing world in section 6.6.

Features of the subject as it is conceived using the boundary concept cannot be deduced from the way in which it is made up, in the way that the features of an organism, such as its mass, its maximum speed and its intellectual capacities, might be deduced from its physical structure and from the arrangement and chemistry of its cells. There is no structure to inspect. The subject as it is conceived using the boundary concept is a construct, used to fulfil a particular role in our understanding of ourselves and of the world. We are free to incorporate in the construct whatever features we choose, subject only to the three constraints of non-

contradiction of our existing knowledge, plausibility and the avoidance of any more than minimal mystery. Such constraints might appear to allow far too much liberty, and even to license an ad hoc construct, the existence of which would prove nothing, precisely because it was ad hoc. But if we found that we had a simple and straightforward construct, which nonetheless did useful work because the subject as so conceived could be seen as engaging in subject origination, then we might regard the construct as worthwhile, and as able to tell us something about ourselves and about our relationship to the world. I shall, however, refrain from incorporating into the boundary concept anything more than is necessary in order for the subject to be seen as engaging in subject origination.

This minimalism is not meant to restrain the construction of fuller concepts of the subject in response to other considerations. One would need something fuller if one were building a theory of the subject that would explain the psychology of our relationships to ourselves and with one another, or a theory that would explain specific philosophically significant features of the subject, such as immunity to error through misidentification, but I shall leave those avenues unexplored. More generally, a concept of the subject with which we could identify would need to be much fuller than the boundary concept. Both the great richness of any acceptable concept of the subject, and the scope for variety in such concepts, are amply demonstrated by two contrasting books, Charles Taylor's *Sources of the Self* and Jerrold Seigel's *The Idea of the Self*. The minimal boundary concept of the subject with which I shall work would be far too thin a concept to be acceptable. But I can set that issue to one side, and develop only a minimal concept, because its use to allow us to attribute subject origination in no way precludes the use of other concepts for other purposes.

The boundary concept must accommodate our sense that we are points of origin, rather than intermediate links in causal chains. It is a desire to accommodate this sense that leads me to present the subject as something that is not within the world. That presentation allows us to see the subject as the source of free but controlled extraneous interventions which are not seen as arising out of the execution of programs that are foisted on the subject. The interventions are seen as emerging from the

boundary, but we cannot see any origins of those interventions. The subject as it is conceived using the boundary concept cannot be seen as having a physical or quasi-physical structure, and cannot be seen as the locus of physical or quasi-physical events within itself. Such features would also be ruled out by the constraint of plausibility. We must not say anything about the subject that would require a quasi-physical world to accommodate its features, alongside the physical world, or anything that would require some new category of physical feature or event within the subject. Furthermore, such features would undermine the deliberate limitation on the status of the originator conception that makes its use merely a way of looking at people who are parts of the physical world, rather than its use affording us a vision of a non-physical world or of strange features of the physical world. Correspondingly, the boundary concept only applies to the whole person, not to the sub-personal mechanisms that are both mental, although easily arguable to be reducible to the physical, and vital to our abilities to perceive, to choose and to act.

The boundary concept must also assign each of us a location in space-time. Location is fundamental to our lives, to what we can perceive and to what we can do. The subject must have a location in the world, even though the subject under the boundary concept is not seen as being within the world. There need be no difficulty here. Each person's boundary is an interior limit of physical space. It is located in physical space, where the physical person is located.

Going beyond the obvious importance of location in space-time for the way in which we live, specifically philosophical concerns also argue for the importance of location. Quassim Cassam argues the case in detail in *Self and World*. He uses a range of arguments to make the case that a unified and self-conscious subject must be embodied. Ingmar Persson uses the importance of embodiment to give an argument against general scepticism about the external world ("Self-Doubt: Why We are not Identical to Things of Any Kind", sections 2 and 3). Brian O'Shaughnessy argues that awareness of one's body is required for the proper use of one's senses (*The Will*, volume 1, part 2). All of these authors argue for embodiment, which is more than mere location, but embodiment implies location.

The subject as it is conceived using the boundary concept can be seen as meeting such embodiment requirements, because the boundary is linked to a physical body. It is not linked as something attached to a physical body like a Cartesian soul, even one that is much more closely integrated than a pilot in a ship, to use Descartes' image in the *Meditations* (Meditation 6, AT 7: 81 (nauta), AT 9: 64 (pilote)). It is linked more in the manner of a feature of the physical body. To revert to visual images, the boundary is like the surface of the body, not something additional but an inseparable feature of the body. But we must not push this image too far. We can only continue to apply the boundary concept so long as the body is alive and has cerebral functions of appropriate kinds, whereas the surface of a physical body exists regardless of the state of animation of the body. We must also remember that use of the boundary concept of the subject does not imply the existence of anything over and above the body. The inseparability of the subject as it is conceived using the boundary concept amounts to the fact that living physical bodies of an appropriate degree of similarity to ourselves, and only they, can be brought under the boundary concept, and that they can be seen as being selves that fall under that concept, not as possessing selves that fall under that concept. These limits to the implications of the use of the boundary concept mean that I need not be embroiled in debates about whether or not people are identical to bodies, or are constituted by bodies, nor need I be embroiled in related debates about the nature of, and the criteria for, personal identity. We would also be free to associate a subject as it was conceived using the boundary concept with two bodies, if Anthony Quinton's fantasy of a double life that was led in two places, with times awake in one place matching times asleep in the other, were a reality ("Spaces and Times", sections 5 and 6).

Finally, the subject as it is conceived using the boundary concept must be seen as a unified subject. It is fundamental to the way in which we regard ourselves and one another that each one of us is seen as a single person. Inner conflicts are resolved before decisions are made. We know what we, and other people, think and do, not what parts of ourselves or of other people think and do. Irresoluble conflicts may be the stuff of novels, but most of life is more straightforward than that, and when there is a conflict, a resolution is sought. Division of the subject is a peculiarly debilitating

mental condition, precisely because our way of life and our societies presuppose integrated selves. (For a full exploration of the philosophical, rather than the practical, significance of a unified rational point of view, see Rovane, *The Bounds of Agency*.)

The motivation for use of the boundary concept of the subject is simply that it allows us to see deliberation with mastery, subject origination and the unity of the subject, without sacrificing location or embodiment. Having said that, use of the boundary concept does not explain how subject origination works. Only the corresponding physical processes remain amenable to explanation, and their explanation remains firmly the province of the natural sciences, which only consider things that are within the world. There is therefore a certain mystery, the creation of which is the price of allowing for subject origination. I shall now outline both the continuing role of the explanations that are offered by the natural sciences, and the extent of the mystery.

The role of scientific explanations is unchanged by the introduction of the boundary concept. There is no suggestion that use of the boundary concept brings something new to the world, something like a soul that is additional to the physical body. Rather, use of the boundary concept is an alternative way of looking at what is already there. Thus the mechanisms of the brain continue to be the condition for consciousness and the source of consistency in our thoughts, words and deeds. Our characters, dispositions, knowledge and actions are consequences of physical states, and not of anything non-physical. We can therefore tackle Schopenhauer's problem without being led to his conclusion. He points out that a person is conscious of himself in two different ways, as a representation and as a will (*The World as Will and Representation*, volume 1, chapter 19). In place of these two modes of awareness, I have the person as he or she appears under the two conceptions, naturalistic and originator. But I do not see what is given under the originator conception as reality, in the way that Schopenhauer sees the Will, because the originator conception merely affords us a way of looking at people. Those people are and remain natural.

The mystery that is generated by use of the boundary concept is not some supernatural world beyond the reach of the sciences. It is a limit to our investigations. The progress of science has accustomed us to the view

that there are no ultimate mysteries. There are plenty of things that we do not yet understand, doubtless including things that we do not yet realize are there to be understood. But we feel that any given mystery is open to scientific investigation, and that there is a good chance that we will in due course solve the mystery. We may come across new mysteries in the process. If, for example, we explain the behaviour of atoms in terms of their constituent particles, we are faced with the question of why those particles behave as they do. But those new mysteries will succumb to science in their turn. At least, there is no reason to think that they will not. A pre-ordained limit to our investigations would be a mystery indeed.

The subject as it is conceived using the boundary concept amounts to such a limit. Nothing that is conceived in a way that makes it amenable to scientific study can, as so conceived, also be seen as engaging in subject origination. If something is conceived in a way that makes it amenable to scientific study, everything that it does is seen either as caused, or as arising merely by chance. It is therefore seen as a locus of events but not as an originator of them. This must include the choices that it makes at the selection stages of deliberations. In order to avoid seeing everything done either as caused or as arising by chance, the subject needs to be conceived in a way that puts it beyond the reach of scientific study. Furthermore, we need to go down that route as soon as we start to think about how we live. We recognize that implicitly seeing ourselves as points of origin is important, both in our individual lives and in our relationships with one another. We then have reason to apply the boundary concept to ourselves, because that allows us to see ourselves as leading the kinds of lives that we know we do lead.

Setting a limit to our investigations is important for another reason. A standard objection to seeing a subject as master over all rational antecedents is that the subject is then seen as lacking any character, so that its decisions must be seen as random. But if we cannot see any origins of the subject's decisions, while we still take it that they do have specific origins, then we cannot say that those decisions are random, any more than we can say that they are not random. This is the reply, promised in section 1.5, to Leibniz and to those who take up that criticism of the image of the queen.

The mystery could be avoided. We could decide not to think about how we lived. But that would greatly impoverish our mental lives. It would eliminate a great deal of the point of literature, of philosophy and of the study of history. But it would not impoverish the natural sciences. The boundary concept accommodates our sense of leading our lives. This is something that would not interest someone who took a scientific view, because scientific views are purged of the specifically human. Science need not feel the pinch of the mystery, because the mystery amounts to placing a territory out of bounds to scientific investigation when science has no interest in that territory anyway.

The mystery could be replaced by something else, equally impure and a touch disturbing when regarded from a scientific stance. Thomas Nagel, in *The View from Nowhere*, plays off subjective and objective views of the world against one another. He accepts that there will always be some tension between the two types of view (ibid., chapter 5, section 6). Living with that tension is an alternative way to accommodate our lives as subjects in an objective world.

The boundary concept and other philosophers

The boundary concept is a minimal concept of the subject. More substantial concepts have been the topic of much philosophical debate. The minimal nature of my concept means that it neither raises nor addresses many of the philosophical issues that surround more substantial concepts. It is, however, still instructive to survey some of those philosophical issues, both in order to underline the minimal nature of the boundary concept, and in order to show that arguments that use it can be conducted without first engaging with those issues.

Gilbert Ryle considered it to be a mistake to seek some given thing that would answer to the pronoun "I", beyond identifying the person one was by name and address. We might think that we needed something extra in order to account for the ways in which we thought and spoke about ourselves, but careful attention to our language would show that this was not so (*The Concept of Mind*, chapter 6, section 6). Ryle's arguments prompt the thought that the boundary concept of the subject might be redundant. But in fact the concept does useful work. The work in question

is not to supply some extra thing that answers to the pronoun "I", in the way that Ryle argued to be unnecessary. Instead it is to allow us to see each person as a point of origin. If we are to make sense of subject origination, we need to identify points of origin. The boundary concept's job is simply to allow us to identify those points.

Before Ryle, David Hume stated that he could never observe himself, but only perceptions (*A Treatise of Human Nature*, book 1, part 4, section 6, page 252). The subject as it is conceived using the boundary concept is likewise unobservable, but does that matter? If we wanted the subject as it was conceived using the boundary concept to play certain roles, such as the role of accommodating introspective perception, that would have important consequences, limiting the range of acceptable theories (Shoemaker, "Introspection and the Self"). But use of the boundary concept does not lead us into such entanglements because we only use the concept in order to pick out points of origin. Incidentally, one issue that is thereby avoided is the issue of whether the subject has intrinsic, as opposed to relational, properties (Shoemaker, "Introspection and the Self", pages 122-123). This chimes nicely with my refusal to give the boundary concept any content beyond what is required for it to perform its function. We might be tempted to develop the boundary concept to do more. Specifically, it might be made the underpinning of care about the future, because that future was the subject's own future. In order for the subject to be seen as fulfilling its duty of care, it would have to be seen to act for its own reasons rather than just letting the future happen. And if the subject were seen merely as an element in the causal network, it would be easy, although not inevitable, to see its future as just happening to it. That would be even further from what we would need to see than the minimally active, because deliberately passive, letting the future happen. The originator conception would prevent us from seeing the subject as passive. Application of the originator conception would in turn imply use of the boundary concept. We might go on to find an answer to Hume in a boundary concept that was developed to accommodate a caring relationship to the future. The subject's projection of itself, as conceived using the boundary concept, into its own future might give it enough substance for it to see itself as real, and to that extent to appear to itself, even though it would still not be an object

of sense perception. Such an approach would lead us into the line of thought of Heidegger in *Being and Time*.

A different approach to the subject is set out in Susan Hurley's book *Consciousness in Action*. The subject is not taken to be a mysterious nexus between perceptual input and behavioural output, something that it would be difficult to treat as a natural part of the world. Both the idea of the given, ready-made perceptual content, and the idea of the giving, actions considered as isolated emanations of a black-box subject, are attacked as myths. A complex picture of personal and sub-personal systems is built up. That picture integrates perception and action. The contents of conscious states are sharply distinguished from the sub-personal vehicles of those contents. On the one hand, the emphasis on action is something that I have in common with Hurley, because many deliberations lead to action. On the other hand, the subject as it is conceived using the boundary concept is very much an input-output subject. It receives information, then as part of a deliberation it makes choices at the selection stage that ultimately lead to actions. The input-output picture that Hurley confines to sub-personal systems is one that I extend to the whole person at the selection stage, although the processing of information that leads up to the selection itself, and the subsequent calculation stage, can be seen as performed by sub-personal systems. There is, however, an important difference between the application of the input-output picture to sub-personal systems and its application to the whole person. Sub-personal systems are embodied in mechanisms within the world, such as systems of cells. We can always look inside the box that represents such a system and see how certain inputs lead to certain outputs. The whole person, as it is conceived using the boundary concept, is not open to such inspection. The boundary that is the person is a recipient of inputs and a point of origin of outputs, and there is nothing to be seen underlying those roles. It is not that any magical or mysterious processing goes on behind the boundary. Processing goes on in the physical world. The boundary plays the logical role of the locus of free choice at the selection stage of deliberation, rather than actually playing any efficient role in the middle of a causal chain. The fact that the subject as a boundary is merely posited in its role, and is not described as a mechanism within the world, is enough to save the subject

from disappearing into the distance as more and more of our cognitive processes are analysed in terms of neural mechanisms.

My input-output picture of the whole person does not set me at odds with Hurley, because our aims are different. Hurley's concept of the subject is a naturalistic one. I seek a concept of the subject that will do justice to our self-conception and to our inner experience. We experience, think and decide. We see ourselves as points of origin. We do appear as input-output selves, both to ourselves and to other people with whom we interact socially, even if a detached observer might find it more satisfactory to see us in a different light, and even though we only need to see our whole selves as input-output selves in relation to the selection stages of our deliberations. Someone who adopted the originator conception of human beings, and who therefore adopted the boundary concept of the subject, could equally well, for other purposes, adopt the naturalistic conception of human beings and adopt Hurley's view in full. Contradiction would be avoided because the conceptions would be different.

Helen Steward, in "Fresh Starts", offers an approach to agent causation, rather than to the subject, that is based on her view that causation is a folk concept that covers a wide variety of relationships. She argues that the concept can easily allow both for talk of substances as causes and for talk of events as causes. If that is accepted, then it becomes possible to avoid the discomfort that is likely to attach to saying that event causation is the norm, but that human agents are a special case in that they engage in substance causation. As with Hurley, my relationship to Steward is determined by the fact that she is looking for something in the world, while I seek no addition to the inventory of the world that is given by the natural sciences. Steward argues that the obstacles that stand in the way of our seeing the world as open to the inclusion of fresh starts, heads of causal chains that are not to be traced back to previous causes, should not be taken to be insuperable but are open to challenge. The doctrine of subject origination does not need to accommodate such fresh starts as real features of the world.

I can likewise avoid certain difficulties that surround mental causation, such as the problem of the implications of the principles of causal closure and of non-overdetermination, and the problem of the

alleged causal inertness of the mental. (For discussions of the first problem see chapters 6 to 8 of Walter and Heckmann (eds.), *Physicalism and Mental Causation*. For the second problem, see chapters 9 to 11 of the same book.) I am not driven to seek mental causes in the world, because I do not seek agential causes in the world. This also allows me to disregard the challenge to folk psychology that is given by Ramsey, Stich and Garon in their paper, "Connectionism, Eliminativism, and the Future of Folk Psychology". The challenge in question is that folk psychology assumes propositional modularity, which is argued to be incompatible with connectionism. The incompatibility has been disputed (Frankish, *Mind and Supermind*, chapter 6). But in any case the alleged incompatibility does not affect my proposal, because the argument for incompatibility relies on incorporation into the doctrine of propositional modularity of the claim that propositional attitudes play a causal role. This is a claim that I do not need to make. Such a causal claim would even be at variance with a view of deliberation as conducted with mastery, and not as the following of a program, if the claim went so far as to allow propositional attitudes to be sufficient causes of outcomes. The claim that propositional attitudes play such a strong causal role may, however, be sustainable when one takes a different view of deliberation.

Finally, E. J. Lowe, in *Personal Agency*, presents a special version of agent causation that is supported by two key arguments. The first argument, that all causation is fundamentally substance causation, eliminates an otherwise awkward contrast between scientifically respectable event causation and mysterious agent causation. The second argument, that there are causally efficacious non-epiphenomenal volitions, gives a way of conferring agential control. Furthermore, our actions are not random. They are performed for reasons, which in Lowe's view not only are not causes, but must not be causes if our actions are to be both free and rational.

Lowe is, like Steward and others, looking for a satisfactory picture that is wholly based on things in the world, although Lowe is clear that important things may fail to feature in a purely physical description of the world. Volitions have real work to do in Lowe's theory, so they must be real. That fact is reflected in a key difference between Lowe's approach and my

own. I am happy to see the physical facts, facts that can only allow for determinism and randomness, as capturing all that can objectively be said about the world. Relatedly, I do not conclude that there need actually be a causally efficacious will, even though my argument that we should see ourselves as deliberating with mastery is motivated by similar concerns to those of Lowe, who says that we must not see ourselves as reasoning automata or as caused to act by our reasons (*Personal Agency*, sections 4.7 and 7.9). Another difference between Lowe and myself is that I locate freedom, as seen but without any claim as to its reality, at the selection stage of deliberation, not in any volition that is the direct impulse to action.

I have a specific concern about Lowe's approach which leads me to think that his proposal is no better than mine, although no worse either. Lowe asserts that control comes from our volitions, and that our actions are not random because they are performed for reasons. He does not offer, and does not want to offer, a physical mechanism of control. But only a physical mechanism, or mental action on the physical, can actually control physical action. Lowe naturally chooses the latter option, but he thereby introduces an element of mystery into his proposal that is no less unnerving than the element of mystery in my proposal. He does give arguments to show that we should not rule out the sort of action of the mental on the physical that he needs, but to say that something is not ruled out is not to argue convincingly in favour of its actuality. Lowe freely admits that there is an incompleteness in his proposal at this point (ibid., section 8.5). My unbacked promissory note on control represents a comparable incompleteness in my proposal, although I at least have the defence that I am not asserting that there actually are open but controlled choices, only that we can see our deliberations as if there were such choices.

In the first three chapters, I have argued that the naturalistic conception of human beings is not adequate to our experience or to our self-conception. There are therefore good reasons to adopt the originator conception. In chapter 4, I shall set out some connections between the concept of rational action, action that is the result of a process of deliberation that we see as rational, and a range of other concepts. Then in chapters 5 and 6, I shall discuss a range of questions in epistemology and in the philosophy of science.

CHAPTER 4

Rational action

In this chapter, I move on from how we see processes of deliberation, to how we evaluate their results. The results are, however, not to be evaluated in isolation. If a choice of action is to be seen as rational, the agent must be seen as having deliberated appropriately. This shift from the inner to the outer is described in section 4.1. I use the concept of rational action as a centrepiece, around which to arrange several other concepts. In section 4.2, I explore the relationships between rational actions and reasons, and survey the attitudes that an observer of a putatively rational action might have toward its rationality. I also touch on the source of our standards of rationality. In section 4.3, I discuss the concept of self-consciousness and its relationship with the concept of action. In section 4.4, I discuss the concept of consciousness, and note the limited extent to which consciousness need be understood for present purposes. In section 4.5, I discuss the concept of choice and the constraints on our attributions of choice.

4.1 From the inner to the outer

The focus so far has been on the process of deliberation. There are benefits in seeing people as deliberating with mastery, and in overlooking the causal closure of the physical. That much concerns seeing the process as non-mechanical. But if we want to attribute rationality, we must look at deliberation in a different way. We must consider the content of the evidence that is available to a subject, the actual weights that the subject attaches to pieces of evidence, and the method of argument that the subject

selects. A move from the inner to the outer, from the view of the process as not proceeding mechanically to consideration of the content of the process and of the result, is appropriate because the conclusions of deliberations, and the relationships between the content of those conclusions and the content of the supposed reasons for them, are what we take to indicate a subject's quality of thought. We do not credit the wild eccentric with a rationality that makes it worth our while to pay any attention to him, even if we regard him as deliberating with mastery.

The move that is intended here is not a complete move from process to result. The rationality that is at stake is not simply a rationality of the conclusions of deliberation, considered in isolation. A conclusion will be considered to be rational if it is related to the starting point in evidence by an appropriate set of weights and an appropriate method of argument. But if a judge of rationality would not have chosen the action, or adopted the belief, that the subject chose or adopted, that would not be sufficient to make the choice of the action or the adoption of the belief irrational. Rational action is neither intensionally nor extensionally equivalent to ideal action. Likewise, rational belief, in the sense that I intend, is neither intensionally nor extensionally equivalent to correct belief, because we may conduct research rationally while still not arriving at the correct answers to our questions. I shall open the discussion in terms of rational action. The rationality of holding beliefs will come to the fore in chapter 5.

Several concepts are in play here. We can only make sense of the concept of rational action if we can make sense of the concepts of a reason and of rationality. Furthermore, only a self-conscious being can decide that he or she will do something. Self-consciousness in turn demands consciousness. And rational actions are actions that are chosen on the basis of reasons, so we must be able to make sense of the concept of choice in order to make sense of the concept of rational action.

These concepts can all be arranged around the concept of rational action. Rational action is something with which we are all familiar. We engage in it constantly. We recognize it in ourselves and in others, whether or not we understand its true constitution. It is a suitable centrepiece for the complex of concepts with which we are concerned, both because of its familiarity and because there are reasonably direct connections between it

and the other members of the complex. These connections are largely connections of presupposition. The other concepts are presupposed by the concept of rational action. If they make no sense, then the concept of rational action will not make sense either.

The concept of rational action is sophisticated, so it would not be the obvious starting point if we were trying to ground the whole complex in other concepts that were generally taken to be less problematic, such as concepts that originated in the natural sciences. But I shall not attempt such an exercise in grounding, whether reductive or otherwise. I also recognize that the philosophical issues that are associated with all of these concepts are large. I shall only offer a sketch map. I shall position my concerns within the wider philosophical landscape, but I shall not attempt to resolve debates in the terrain that surrounds my own.

4.2 Rational action and reasons

The most directly connected concept is that of a reason. A rational action is an action that is performed for reasons. But it is not only that. An action may be seen as performed for reasons, but not be seen as rational because the reasons were not good enough to justify performance of the action.

If someone performs an action for reasons and someone else then asks why he did what he did, he will have something to say to her, something that will go beyond the absence of reasons not to have performed the action, even though going for a walk "for no particular reason" is not to be condemned as irrational. What he can say will amount to the reasons he had, and the questioner will often accept that they were sufficient reasons. If the questioner does not accept the putative reasons that are given, and replies that they were not really reasons, or that they were insufficient, we can expect a dialogue to ensue. The agent may offer an explanation of why the putative reasons were reasons, or sufficient reasons, or he may offer additional reasons. He may also explain the description under which he brought the action at the time, lest there be some misunderstanding there. The questioner may well grasp the point eventually, perhaps concluding that she would not herself have performed that action for those reasons, but

that the action was nonetheless rational. Any rationality that is attributed will not be rationality by the agent's lights, but rationality by the questioner's lights. There is a sequence of four attitudes, all of which appear to be possible but the second one of which actually makes no sense.

The first attitude is that the action was not rational. This is the likely attitude if the questioner considers that the agent's weighting of evidence, or his chosen method of argument, was unacceptably peculiar. That is, it is the likely attitude toward the truly bizarre eccentric. It is also the likely attitude when the questioner recognizes that the agent went through a process that resembled deliberation, but considers that reasons did not really enter into that process because the putative reasons would not have been able to play their required roles in a genuine process of deliberation. This could be because the putative reasons were, in the questioner's eyes, irrelevant to the decision that was to be taken. Instead, there was what she would regard as some non-rational process, with the putative reasons being seen by her as made up independently. Rationalization is not rationality. The questioner could also see rationalization when she thought that the reasons supplied were perfectly relevant reasons, but when she thought that they had been misused in the supposed process of deliberation, because either the logical structure of the process, or the relationship between that structure and the reasons, was unsatisfactory. In such cases, it would however be difficult for the questioner to spot that something was amiss. As will be apparent from this paragraph, a decision that merely happens to accord with what we would regard as good reasons is not rational in the sense that I intend. The process is important, as well as its result.

The second attitude of the questioner is the one that would be needed for an attribution of rationality by the agent's lights that was not an attribution of rationality by the questioner's lights. This attitude is that the action was rational by the agent's standards, which the questioner could not grasp. (If the questioner could grasp, but could not accept, the agent's standards, the first attitude would be appropriate.) This attitude makes no sense because rationality is not something highly technical. The concept of rationality plays a role in the everyday understanding of people on which we rely in our social interactions. The technicality of a process of deliberation is no bar to a grasp of the rationality of the ensuing action,

even though it makes perfect sense for someone to say, "I am sure that he had his reasons, but I cannot understand them because they are too technical for me". The agent might, for example, have carried out a business transaction using complicated legal formalities, and the questioner might be unable to understand the detailed reasons for the formalities. But there would be something that the agent could say, and that the questioner could grasp. The agent might say that the point of the formalities was to reduce a tax liability. That simplified account would allow the questioner to see the action as rational. There are some apparent exceptions. For example, a mathematician might set out the steps in some complicated work that he was doing in order to establish a particular result, and lay people might be unable to comprehend the result, nor would it have any practical application that might strike lay people as sufficient to motivate the work. In such cases, it might seem that a questioner would have to acknowledge the incomprehensible rationality of the actions. But such exceptions are rare, and are liable to turn out only to be apparent exceptions. The mathematician could certainly explain to a questioner the notion of solving a complicated problem by roundabout means, the notion of a proof on which one could rely and the notion of adding to our knowledge out of interest, or because the addition might contribute to the acquisition of further knowledge, even if the mathematician could not explain the content of his work to lay people.

Another reason why there is no space for attributions of incomprehensible rationality is that if a questioner attributes rationality to an action, she implicitly claims that she comprehends the agent's reasoning sufficiently to have at least a rough sense of how he might have reached his decision, even if she would never actually think like he does. Two arguments that are closely related to each other can be used to support the view that the questioner must have at least some comprehension of how the agent reached his decision, if she is to make any attribution of rationality. The first argument is that a state of the world cannot contribute to making an action rational for the agent, that is, serve as an internal reason for the action, unless the agent has phenomenal consciousness of that state of the world (Eilan, "Perceptual Intentionality, Attention and Consciousness"). If that is correct, then a questioner will be unable to

appreciate that an agent acted rationally, unless the questioner has at least some grasp of the agent's form of awareness of the world. It might be enough for the questioner to have only a very limited grasp. If we started with human beings and worked outward to stranger and stranger aliens, we might reach other limits to comprehensibility before reaching this one. Nonetheless, the argument can be used to support the view that an ability to grasp the processes of thought of others to some degree is important. The second argument is that a questioner can only properly understand a process of choosing an action as rational if she can appreciate what matters to the agent. She must therefore be able to grasp the emotions of the agent, at least to some extent. Those emotions cannot be grasped in a detached way, nor can one have an adequate detached grasp of concepts such as "the dangerous". From a wholly detached point of view, a questioner could only grasp "the dangerous in the view of the agent", not "the dangerous" as it was from the agent's point of view (compare Goldie, "Emotions, feelings and intentionality", pages 244-246). Such considerations contrast with my remarks in section 3.3, about the lack of need for qualitatively similar inner experience in order to enter into another deliberator's head, but there is no contradiction. In section 3.3, the task was merely to grasp a sequence of thoughts as a process of deliberation. Here, the task is to attribute rationality to the deliberator's conclusion. That attribution could be withheld, even if one saw the deliberation as following an intelligible course.

The exclusion of attributions of incomprehensible rationality does not exclude an acknowledgement that others might have standards that played a comparable role in their lives to the role that our standards of rationality play in our lives, even though those other standards might be too alien to make any sense to us. But our inability to make sense of those standards would prevent us from applying those standards to particular actions. The best we could say about the actions of such mysterious beings would be that a given action might, for all we knew, conform to standards that played a normative role for them that was comparable to the normative role that standards of rationality played for us. That would also be the worst that we could say about their actions. We could not condemn the actions as irrational, any more than we could honour them as rational.

Global incomprehension would require us not to reach a conclusion on the status of the actions.

The third attitude of the questioner is that the action was rational, even though the questioner might not herself have performed it on the strength of the reasons given. The questioner might not have attached the same weights to the pieces of evidence as the agent did, or she might have preferred a different method of argument, but she could still see that the agent might reasonably have thought as he did. Alternatively, the questioner might not grasp the reasons in all of their technical detail. The fourth attitude is that the action was rational, and that the questioner would have done the same thing on the strength of the reasons given. These last two attitudes both make perfect sense.

The exclusion of attributions of incomprehensible rationality goes hand in hand with the view that rationality is in practice rationality for us. We have no access to standards other than our own standards. This is, incidentally, the reason why we can solve the coincidence problem, the problem of explaining why it is that intentional reasons-based explanations of conduct and non-intentional physical explanations march in step. (See Schlosser, *The Metaphysics of Agency*, pages 123-126, for a recent analysis of the problem.) The standards of rationality that we have created govern our identification of some pieces of evidence, and not others, as potential reasons for given actions. They also govern our identification of some constructions that use evidence, and not other constructions, as rational justifications for given actions. If agents use acceptable constructions that use appropriate evidence to reach their decisions on what to do, those constructions amount to acceptable reasons-based explanations of their actions. We have chosen our standards because those standards allow us to make sense of human beings. If we had chosen standards that led to the identification of types of reasons-based explanation that did not march in step with the physical progress of the world, we would have found that our reasons-based explanations did not work. If we had tried to use explanations predictively, people would not have acted in the ways that the explanations would have led us to expect. If we had tried to use explanations to explain actions after the event, either we would not have been able to find explanations for actions that were performed, or we

would only have been able to find them because our standards allowed such a great variety of types of explanation that an explanation could have been found to fit almost anything. Standards that were so lax would barely have amounted to standards at all, and they would have been useless in identifying types of explanation that could have been used in prediction. Given that we want standards that steer us toward at least some success in prediction, it is not surprising that we have chosen standards that lead us to identify explanations that work, given the physical facts about human beings, about their way of life and about their environment. So it is not too difficult to find alignment between reasons-based and physical explanations of human conduct. It would probably have been a great deal more difficult if the standards had not been chosen by us and tailored to our way of life, but were universal, binding on all rational beings and purely discovered, rather than being at least partly invented. Finally, the fact that it is possible for there to be pairs of reasons-based and physical explanations that march in step is no great mystery. Each human brain is both sufficiently complex, and sufficiently consistent in its physical operation.

Judging rationality

There is no available standard of rationality that we can be confident is independent of culture. But the risk that standards of rationality are culture-relative does not mean that there is a free-for-all in attributions of rationality. While anyone may claim that his actions, or specified actions of another person, are rational because he sees them as such, there is no reason for us to accept such claims unless the actions satisfy our own standards of rationality.

In order to have a concept of rationality that is sufficiently stable and widely enough used to be useful, we need the right sort of standards to govern the attribution of rationality to actions, neither too strict nor too lax. There are three points to consider. First, the range of propositions that a judge of rationality would in principle accept as reasons for a given action must be appropriate. Second, the judge must require the agent to have an appropriate degree of justification for belief in the propositions that are cited as reasons. It would be appropriate to deny that an action was

rational if it was chosen on the basis of a supposed reason that the agent had too little justification for believing, but the standard of justification should not be set too high. Third, the judge must apply appropriate standards to the ways in which the propositions that are cited as reasons are used in choosing actions. The judge should require a degree of logic, but should not be too strict on this point. On all three points, the judge should apply appropriate standards, not merely quirky personal standards. Before turning to the source of appropriate standards, I shall make two incidental points. The first concerns the distinction between internal and external reasons, and the second is specific to external reasons.

The first incidental point is that we need not worry about the distinction between internal and external reasons, because we are concerned with judgements as to whether agents select their actions appropriately, given the data that are available to them, including data to the effect that further relevant data can be obtained and considered, and those further data themselves. We can limit ourselves to internal reasons, propositions that can motivate the agent given his existing set of motivations and assuming, if necessary, that he is better informed about the state of the world, including himself, than is in fact the case. (The limitation to data that the agent has or that he can obtain imposes an additional restriction, because there may be unavailable data that would motivate the agent if only he could obtain them.) We need not concern ourselves with external reasons, propositions that are reasons but that could not motivate the agent. But once we understand internal reasons, we will be able to understand external reasons on the same pattern, assuming that there is any such thing as an external reason. It is true that judgements of rationality will reflect the propositions that the agent could in fact consider as possible reasons. But in the course of consideration of an action's rationality, the propositions will simply be potential reasons for the action. It will be taken for granted that they could be reasons for the agent, but that fact about them will not be reflected in the course of the consideration.

It might seem that there was a special situation in which the requirement for the agent to have justification for his belief in the propositions that were putative reasons would prevent this easy move to

an understanding of external reasons. This would be when the reasons were external because the agent could have no access to estimates of the probabilities of the truth of the propositions that were any more than baseless guesses, on account of something more fundamental than a lack of resources to expend on gathering data. But there is no problem here. We can ignore the issue of justification when considering external reasons, because we can simply take the propositions concerned to be true, and then consider whether they would be reasons for given actions if the agent did justifiably believe them. Even the existence of external reasons that were external because the agent could not understand the propositions that gave those reasons would not debar us from having a reasonably general understanding of external reasons, even though their existence might make it difficult to arrive at a completely general understanding. The agent's inability to understand propositions might reflect limited mental capacity. We could then suppose the deficiency to be made good without changing what would or would not be a rational choice of action, except in the proportionately rare circumstances where mental capacity would significantly affect the range of reasonable actions. Examples are when possible actions would include applying to a university, so that intellectual capacity mattered, and when they would include having children, so that emotional capacity mattered. We might save the day even in cases like these by supposing the agent's mental capacities to be enhanced for the purposes of deliberation, while also supposing the agent to take into account that his capacities would be reduced to their actual level after the deliberation had been completed.

The second incidental point is that the requirement for an agent to make appropriate use of propositions that he takes to be reasons would be no bar to the extension of our understanding to external reasons. External reasons could be imagined to sit within an argument for a given action that a different agent might have constructed and that could be appraised on its own merits, independently of what sorts of argument the agent might himself construct. It is worth noting that a special form of externalism lurks here. Once we get beyond relatively simple relationships between a desire for a certain outcome, a belief that a certain action will bring it about and a decision to act, and move on to more complex relationships

between beliefs and actions, an agent may be aware of reasons for an action and may have an intuitive sense that they are reasons for that action, but may be unable to articulate the connection between the reasons and the action in the way that someone with a better grasp of logic could articulate it. This would be an externalism of rationality rather than of reasons. It could be dissolved by an assumption that an agent might have been intellectually more capable than was in fact the case, a parallel move to the standard assumption that he might have been better informed than was in fact the case. That move should be uncontentious except in the cases, such as application to a university, where intellectual capacity would itself affect the rationality of possible actions. It would not go as far as the more contentious move of identifying the use of certain value-inducing patterns of thought as a necessary condition of practical rationality, as in Thomas Nagel's book, *The Possibility of Altruism*.

The source of standards of rationality

We can now reflect on the source of appropriate standards for a judge of the rationality of an action to impose. There are three points that a judge must decide, whether propositions are of the right nature to be reasons for a given action, whether the agent has sufficient justification for believing those propositions, and whether the agent uses them in an appropriate way in choosing the action. The only source of standards that is in practice available is the same for all three points. It is the society within which the judge lives and works. This does not have to mean the whole population of her town, her country or the world. In an academic discipline, where the action to be judged is likely to be the action of conducting research in a particular way in order to answer a difficult question, it means the society of experts in that discipline. In a practical art, such as making violins or flying aeroplanes, it means the society of skilled practitioners. In everyday life, on the other hand, it may well mean the general population.

It is clear that the only people to whom we can turn in the search for standards are the people around us, although they do include our ancestors, whose wisdom has helped to form our current culture. But this lack of an accessible external standard need not be depressing. It does not mean that we must be subservient to the majority, or that we must be governed by the

dead hand of the past. It is perfectly legitimate for a judge of rationality to reject generally accepted standards in relation to the three points on which she must decide, although the majority may then decline her judgements.

We can now move on to the concepts that underpin our understanding of the agent's reasoning before he takes action, the concepts of self-consciousness, of consciousness and of choice.

4.3 Self-consciousness

Self-consciousness is essential because the rational subject must think, "I shall do this", and not merely, "This will be done". The intrinsic link between rational action and the sense of oneself has been set out by John Perry in "The Problem of the Essential Indexical". If I am to act on the basis of reasons that I recognize as such, I must think of myself as myself. Perry uses this point to argue against a traditional proposition-based approach to at least some of our beliefs, but the point stands independently of that use. Lynn Rudder Baker develops the point. She argues that a strong form of self-awareness is needed in order to account for the strength of motivation that such awareness can support, as when Oedipus blinded himself on realizing what he himself had done (*Persons and Bodies*, pages 76-79). Baker, in line with Perry but giving a fuller analysis, distinguishes between a strong form of first-person awareness that involves conceptualizing oneself as oneself, and a weak form that merely involves situating oneself perspectivally (ibid., pages 59-69).

The grounding for the sense of self that I shall offer here is a logical grounding, not a psychological or neurological one. It is also a grounding for the sense of self, rather than a grounding for the self that is sensed. Our inner sense of action plays a critical role. This is one truth in Nietzsche's claim that one does not say "I". Instead one does "I" (*Also Sprach Zarathustra*, part 1, section 4, "Von den Verächtern des Leibes"). It is action, not the observation of oneself or of others, that grounds a sense of the self of which one is conscious as oneself, and hence that grounds self-consciousness. I shall take self-consciousness to require an object that one correctly regards as oneself, so that when there is no such object, there is

no self-consciousness. There would certainly be none of a sort that would allow rational action. Indeed, rational action requires the correct identification of a physical object, rather than an object of any type, as oneself.

Limitations on our action have an important preliminary role to play. We cannot instantaneously jump around the Universe in order to perceive, or act on, different things. Such limitations give a practical sense of what it is to be a physical being (although not yet a sense of what it is for oneself to be a physical being), within a world of physical beings and subject to the world's physical laws, including the laws that require everyday forms of influence to be local, rather than at a distance. We acquire that sense simply by acting, and finding that change can only be effected locally. There is no circularity in using the concept of action at this point in the argument, because it need only be action in general, not action on the basis of reasons. We have a direct capacity to act which means that we can act without relying on a conscious sense of self. Indeed, non-human animals presumably do precisely that. In parallel with this, even an agent who lacks self-consciousness can have a sense of the here and now, acquired by reaching out and touching the world and by perceiving the world and changes in it that follow his actions, although he must lack a sense of himself being here present.

Armed with a practical sense that effective action is generally local, the agent can work out the practical ways in which it is possible to effect change. He can articulate the principle that the person who is here present is the one who must put in some effort in order to effect change here and now. That allows him to conclude that some things that happen here and now are consequences of the actions of the person who is here present. They are not consequences of changes elsewhere that magically transmit their influence over some distance. That allows an explicit identification of the body here present as the agent. All that remains is to identify the agent as oneself. Doing that will give the indexical that is needed. It will do so by linking a train of thought to a physical body that acts. It would, however, be a mistake of a Cartesian flavour to think of the train of thought as belonging to a mental "I" that was waiting to be hitched to a physical body. It would be better to say that we could only speak of some thoughts,

and that we were not entitled to give them a self-conscious owner until we gave them a physical owner (compare Lichtenberg's comment on the Cogito, Sudelbücher K.76: *Schriften und Briefe*, volume 2, page 412). In order to avoid making such a mistake, we must see the subject as acquiring self-consciousness and identifying himself with the agent here present at the same instant.

A subject can appreciate that the agent who is here present is himself, giving the necessary identification, because he is directly aware of acting. A subject recognizes an action as his own because he directly initiates it, and because he senses it as an exercise of a personal power to produce change. As above, there is no circularity in using the concept of action at this point in the argument, because it need only be action in general, not action on the basis of reasons. It is recognized as action in the world, rather than an internal fantasy, because the agent perceives the resulting change in the world. A fuller argument that we can avoid circularity is given by Lucy O'Brien (*Self-Knowing Agents*, page 88). She gives a fundamental role to the agent's awareness of his actions, claiming that this awareness is an awareness by means of the production of actions, rather than by means of the reception of data, and that this awareness does not presuppose a capacity for first-personal reference and thought. The awareness by production that O'Brien identifies can also be invoked earlier in my argument. It allows the subject to identify actions and an agent from the start. He is not limited to thinking in terms of changes in the observed arrangement of limbs that are correlated with changes in the observed arrangement of objects with which those limbs come into contact. I shall return to O'Brien's work in section 6.4, in order to show how we can avoid illegitimate circular reliance on the concept of causation.

As Perry has shown, the indexical is essential to action on the basis of reasons that the subject recognizes as reasons. I have just indicated how action in general, rather than action on the basis of reasons that the subject recognizes as such, can give us a route to the indexical. I shall now argue that we have to rely on action and on our sense of action.

It is hard to see what else, apart from action, could ground the identification of a particular being in the world as oneself that is needed to support the concept of rational action. (Compare the importance that can

be attached to a putatively conscious machine's possession of a capacity for action. See Aleksander, *The World in My Mind, My Mind in the World*, pages 46-47 and 177.) Observation of the beings in the world from some neutral standpoint could not do the job. Observation of which being in the world responded without mediation to one's decisions to act could do the job, but that would involve both action and a sense of action. A sense of action would be required in order to identify what was going on as decisions to act. Identification of the being that was located at the point of origin of one's own spatio-temporal frame of reference, the point with co-ordinates (0, 0, 0, now), as oneself might be enough, but this would not get us away from the need for action. It would not do so because a subject would need a sense of his having a spatio-temporal location, in order to make sense of the idea of his being located at some given co-ordinates. If he were not to use the route to the indexical that was outlined above, he could only get that sense by performing some action, if only the minimal action of processing sensory data. The mere passive reception of data, as impressions on a wax tablet, would not even be enough to give a sense of one's being located in time. That sense could only arise out of the tagging of data with indicators of their order of reception, and a review of the serial order that was thus constructed. (I take it that time exists, in a sense that is strong enough to give an order of reception, independently of our having a sense of our location in time. The spatio-temporal world is conceptually, as well as historically, prior to our insertion into that world, as will be argued in section 6.6.) That tagging and review would amount to active processing. Furthermore, it would only give the subject a conscious sense of his having a location in time if he were conscious of the tagging as an action, because he could only understand what the order of the tags signified if he was aware of how they had come to be assigned. The subject could most easily obtain a conscious sense of his having a location in space by changing his position and seeing how the available data changed, as various objects came into view or fell out of view. A sense of having a spatial location might also be obtained by noticing specific features of the data such as the clarity of edges, or the interruption of some edges by others when one object partially occluded another, and interpreting those features as consequences of the values of newly-invented

variables that were called "direction" and "distance". But those variables could only be interpreted as "direction from me" and "distance from me" if the subject also moved, or at least changed the orientation of his sensory organs, and saw how the observed features changed as he did so. Movements and changes in orientation would constitute actions, and their effects on sensory data would only serve to give the subject a conscious sense of his having a location in space if he was aware of them as his actions, so that he understood the changes in sensory data as consequences of changes in his own location, or in the orientation of his own sensory organs. As Gareth Evans put it, in the context of a more elaborate argument and with the emphasis on a wider class of actions than those that directly change perceptions, "An egocentric space can exist only for an animal in which a complex network of connections exists between perceptual input and behavioural output" (*The Varieties of Reference*, page 154).

The argument that a subject would need to be aware of movements or of changes in orientation, or of processes of tagging, as actions, in order to get a conscious sense of his having a location in space-time, might not appear to work. Would it not be enough for him to be aware that a movement, a change in orientation or a process of tagging had occurred, without having a sense that its execution was an action? This objection can be dismissed. The reason is that the task is not just to give the subject a sense of space and time, but to give the subject a sense of his having a spatio-temporal location, so that he can identify the being that is here present as himself. (The subject must be seen as acquiring self-consciousness, acquiring a sense of his having a spatio-temporal location and identifying the being that is here present as himself, all at the same moment.) A description of a process that took place, and an output that included both sensory data and a statement that the data had been collected from a place at 54 degrees 43 minutes north, 20 degrees 31 minutes east, and 5 in the morning local time on 22 April 2009, would not be enough. The vital extra information would be that the location from which the data had been collected was the subject's location. Once a subject was in the habit of identifying a being in the world as himself, it would be enough to give the subject some co-ordinates and to tell him that they were his co-ordinates. But that would only be enough if the subject already had

a sense of himself as a spatio-temporal being. In order to gain that sense in the first place, he would both have to perform actions, and have to be aware of them through his sense of action. He would also have to note their effects, either in changing current sensory data (for spatial location) or in regimenting historical sensory data (for temporal location).

So there is an intimate relationship between the concept of action and the concept of self-consciousness, whichever route to the essential indexical we use. We need to act, and we need to be aware of acting, in order to have self-consciousness of a type that will allow us to progress to rational action. In the other direction, we do not need self-consciousness in order to act, but we do need self-consciousness in order to see what we do as our own actions. Only that can make it possible for us to see ourselves as considering what we should do in the light of evidence, and only thus can we see ourselves as acting rationally.

With a sense of oneself well-grounded, one can attribute self-consciousness to others, on the perfectly reasonable ground that the evidence of their appearance and conduct strongly suggests that they are like oneself. I do not claim that this is, as a matter of psychology, the order of procedure. As children, we may well develop our senses of ourselves in parallel with our senses of others. I also do not claim that our attribution of self-consciousness to others is in practice the affirmation of a proposition to that effect, rather than its being a deep feature of our way of life. As Wittgenstein remarked, "My attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul" (*Philosophical Investigations*, part 2, section 4, page 152). But once our senses of ourselves and of others have been developed, we can locate their logical foundations in action and in the similarity of others to ourselves. This is, however, subject to the point that the development of a full concept of the self, as distinct from the minimal concept that is necessary to support the concept of rational action, would require a great deal more.

4.4 Consciousness

A creature could not have self-consciousness that was sufficient to allow

rational action if it did not have consciousness of objects in the world. The reason is that it would have no sense of acting on the world, not that it would lack a sufficiently complex brain. It would be perfectly possible for a creature with a complex brain to lack a sense of action and therefore lack self-consciousness of the required sort, either because it was cut off from the world and received no sensory inputs that correctly indicated the effects of its conduct, or because it had those inputs but could not bring the data to consciousness in the sense of being aware of its own experience.

To start with the lack of correct inputs, a brain in a vat, supplied with inputs that were like those that an embodied brain would receive from perception and from action, could come to be in a state that had the phenomenology of self-consciousness. That state might be regarded as one of self-consciousness. But it could not be self-consciousness of the sort that would allow rational action, because of the lack of a physical object that the creature correctly identified as itself. The apparent body would not really be there. Even if the entire body existed, in the vat, the creature would still fail correctly to identify an object as itself because all of the creature's descriptions of the body's situation would be radically false. The creature would pick out a body that was supposedly walking down the street, when no such body would respond without mediation to the creature's decisions to act. (The body in the vat might be immediately responsive, if the vat allowed space for thrashing about and if the body was not paralysed, but it would not be walking down the street.) That is, I do not see a mere identification of some physical object or other as oneself, an identification that might appear to be made correct by the existence of neural pathways connecting brain to body, as sufficient to meet the standard of correctly identifying a physical object as oneself that must be met in order to allow for rational action. At the other extreme, if a brain controlled a body that was acting in the everyday world and that was receiving sensory inputs in the ordinary way, the whole creature could have self-consciousness of the required sort, even if the brain was stored somewhere far away from the rest of the body, and communicated with the rest of the body by radio. But it would be important that the body was the brain's sole, or overwhelmingly main, point of contact with the external world. That would be necessary in order to ensure that the creature was not merely playing at being where

the body was, as in Evans's example of a remotely controlled submarine (*The Varieties of Reference*, pages 164-167).

Turning to the importance of consciousness of data, the most natural way to characterize the bringing of data to consciousness would be to speak of the data being placed in the schema, "I perceive ...". But that would give a false impression of a conceptual circle. The sense of self would not only be seen, correctly, as grounded in action and in observation of the results. It would also appear to be essential to reflection on what one observed, reflection that was in turn necessary to ground self-consciousness because without that reflection, one could not realize that one's actions had consequences in the world and were therefore actions of the self located in the world. In fact, there is no circle here. The data that are supplied by observation do need to be brought to consciousness, but this can be done without using the indexical at that point. One can move a book and say, "The book was on the desk but now it is on the shelf, a change that is a consequence of the action that has just been performed". One does not need to say, "I saw the book on the desk and now I see it on the shelf".

Only some aspects of the concept of consciousness in general are philosophically problematic. It is not problematic that an organism should receive data from its environment and change its internal state or its conduct accordingly. It is not even problematic that an organism should do so intelligently, recognizing dangers and ways to satisfy its needs so as to avoid the former and take advantage of the latter, although that clearly requires considerable sophistication. It is not easy to build robots with comparable skills that they can exercise outside artificially simple environments. The philosophical difficulties really start when we reach what David Chalmers has called the hard problem of consciousness. This problem is to explain why we have an experienced inner life, a life full of qualia, and to explain what our experience is (Chalmers, "The Hard Problem of Consciousness").

Not all philosophers agree that there is a problem here. Daniel Dennett argues that the very existence of the hard problem, over and above a collection of easier problems, is an illusion (*Sweet Dreams*, chapter 3). Daniel Hutto considers the hard problem to be both insoluble and a consequence of a mistaken approach to the topic ("Impossible problems

and careful expositions: Reply to Myin and De Nul", page 50). But when we reflect on the experience of seeing a landscape, of hearing music or of rubbing our fingertips on velvet, there does appear to be something special going on. It is something that seems to be impossible to explain by reference to the interaction of material components, as Leibniz noted with his image of a thinking and feeling being that was enlarged to the size of a mill, so that we could walk in and observe its mechanism (*Monadology*, paragraph 17). But we should not just accept Leibniz's view. Recent advances in neurology, and particularly the identification of processes by which disparate pieces of neural activity may be synthesized into a single experience, disclose possibilities that he could not have considered.

Fortunately, the hard problem does not need to be solved in order to have a grasp of the concept of consciousness that is sufficient for present purposes. Action must be taken on the basis of the lessons of experience, and in the light of current empirical data, if it is to be rational. But the lessons of experience could be recorded, and the data could be taken in, in ways that would allow them to play their roles even if there were no qualia. An account of rational action need not depend on an account of qualia as they might attach to experience and to data, even if our experience is in fact infused with qualia. Rational action also requires self-consciousness, but that too should be achievable in the absence of qualia. It is true that in deliberating, in deciding and in acting, we have particular feelings that are special types of quale. No doubt the presence of such feelings is important in our psychological development, in our coming to acquire senses of ourselves, of others and of our place in the world. But those feelings should be dispensable. If "I shall do this" could make sense in the mouth of a zombie, a being who was devoid of all qualia, including those that were associated with deliberation, with decision and with action, then an account of qualia would not be needed in order to give an account of rational action. I see no specific reason to doubt that "I shall do this" could make sense in the mouth of a zombie, both to the zombie himself and to we who observed him, so long as the zombie could be directly aware of a movement as an action, the awareness being by means of the movement's production. That much would be needed in order for the zombie to have a concept of action, and hence to have self-consciousness as a being in the

world. It would therefore be needed in order for us to attribute self-consciousness to the zombie, because we would not wish to attribute self-consciousness to a being whom we did not see as himself holding that he had self-awareness. It would therefore be needed in order for us to see the zombie as engaging in rational action. But a zombie's awareness that an action was his own could be supplied if his sub-personal mechanisms attached "action" tags to certain movements, tags which would create direct awareness by means of the production of those movements. It seems unlikely that such tags would need to bring any qualia in their train. It would, however, be important for the tagging to be done by sub-personal mechanisms that could simply tag movements without considering the significance of the notion of action. The tagging should merely have its effect on the whole zombie. If the tagging were done by the whole zombie, circularity would threaten.

The senses of choice and of action that zombies had would differ from our senses because our rich inner experience would not be available to them. Furthermore, our interactions with zombies would not be like our interactions with human beings. Despite these differences, the central concepts of self-consciousness and of rational action would still be applicable to beings for whom the hard problem of consciousness did not arise. The ability to apply those concepts should therefore be independent of possession of a solution to the hard problem, even when the concepts are to be applied to beings for whom the hard problem does arise. The validity of this extension to beings with qualia would depend on the absence of relevant differences between the contents of the concepts of self-consciousness and of rational action when they were applied to beings who had qualia, and their contents when they were applied to beings who did not have qualia. But given that the concern is with the logical, rather than the psychological, content of the two concepts, I see no reason to think that there would be such relevant differences. Even if the addition of qualia did lead to relevant differences in content, not all useful results would necessarily be lost. The logical relationship between the two concepts that existed in the absence of qualia could still be available to us in their presence, even if we lacked a full understanding of qualia, so long as the presence of qualia made parallel differences to the contents of both

concepts, allowing their relationship to be preserved. An analogy would be that if $y = 3x$, this relationship remains the same, and is accessible, even if we are only given that $y + 7 = 3x + 7$.

These conclusions are subject to three caveats and a limitation. The first caveat is that it might not be possible to tag certain movements as actions in a neutral way, without any qualia, and for those tags still to have the necessary significance to the subject. The question is that of whether "action" can be a qualia-free tag. The answer is not clear. But I draw comfort from the distinction between knowledge of qualia and knowledge of how, for example, red things look. That distinction can be used to ascribe a better grasp of sensory experience to zombies than one might expect (Raffman, "Even Zombies Can Be Surprised"). If zombies could have a grasp of sensory experience that allowed them to know how red things looked, it would be implausible to deny them a direct awareness of actions. The second caveat is that there are arguments that phenomenal consciousness plays a role in our economy of thought that is essential to rational action (Eilan, "Perceptual Intentionality, Attention and Consciousness"). If any required phenomenal consciousness were of a type that was beyond zombies, rather than being of a type that Raffman's arguments would allow them to have, they might not be able to act rationally. The third caveat is that zombies might not be possible, a question on which philosophical debate continues. If zombies were not possible, it would be risky to argue to the possibility of something else on the basis of an assumption, for the sake of argument, that they were possible.

The limitation is that our sense of making decisions is not a sense of purely abstract, dispassionate choice. Options are coloured by our feelings and by our emotions. Feelings and emotions have their effects on our deliberations through the qualia of felt urges to do some things rather than others. The prospect of one action may give the agent a pleasant feeling, while the prospect of a different action may give him an unpleasant feeling. It is arguable that feelings and emotions operate through qualia essentially. In that case, algorithms that operated without qualia, but that happened always to lead to the same results, would not be adequate substitutes. So it is not clear that zombies could be guided in the making of their choices by anything that could be said to correspond to our feelings and our emotions.

To that extent, they would be denied full access to our sense of choice of action on the basis of reasons, and hence to our sense of rational action. A consequence is that a zombie might well not be able to attribute rationality to our actions, because such an attribution would rely on the attributor's comprehension of the agent's process of thought. In order to have that comprehension, one would need to have at least some sense of what it would be like to think like the agent. A zombie might only be able to attribute rationality to the actions of zombies who were sufficiently similar to himself. Likewise, we can only attribute rationality to the actions of beings who think in ways that are sufficiently similar to our own ways of thought, and the possession of human-like qualia is an important element in our similarity to others. It would be a mistake to assume that we could always regard zombies who were like us, save for their lack of qualia, simply as cut-down human beings, so that we could always appraise their actions for rationality simply by leaving out the influence of feelings and substituting unemotional algorithms. Sometimes, the less than human is the incomprehensibly inhuman. None of this means that we need a full philosophical understanding of qualia in order to understand the concept of rational action. We may only need to have the qualia, in order to use the concept in relation to other beings who have qualia. Either radically different qualia, or a radical lack of qualia, may block the practical application of the concept. Likewise, as will emerge in the next section, subjective similarity of experience is needed in order to make attributions of choice, although no particular type of experience is needed merely in order to grasp the concept of choice.

4.5 Choice

The final concept to consider is that of choice. We can only see an action as rational if we see the agent as having chosen the action for reasons. We must therefore understand the concept of choice, in order to understand the concept of rational action. It is tempting to say that this requires us to see that the agent could have decided otherwise, either deciding on another action or at the very least deciding to refrain from the action. But then we

run into the fact that the agent is a physical being, and that the physical world offers only determinism and randomness, leaving no room for free but controlled choice. In order to make progress, we need to understand precisely what concept of choice is needed to underpin the concept of rational action.

Consider a mathematician who sees how to prove a result that has previously only been conjectured. She writes out the proof. All of her thoughts and actions may be determined, in the sense that the prior positions and states of elementary particles, and the laws of nature, may ensure that certain neurons fire and that her hand picks up a pen and then moves the pen in certain ways. Her action in writing any given line of the proof is, however, rational, on two levels. On the first level, she wishes to write out the proof in order to enlarge our mathematical knowledge. On the second level, the contents of each line are not arbitrary. Either the proposition that is expressed by a line will follow from what has been expressed in previous lines, or it will be an axiom or an established result. The two levels are not independent. The mathematician will only write out something that counts as a proof, and that enlarges our mathematical knowledge, if the contents of each line are allowed by the rules of proof. But however we may analyse these levels of rationality, we can say that the writing of each line is a rational action. Each line is written because it contributes to the overall goal, and it says what it does because its contents both comply with the rules of proof and take us closer to reaching the demonstrandum. Having said that, there may be no rational alternative. There may be only one way to prove the result, or only one way that is feasible given the current state of mathematical knowledge.

The point of this example is that an action may be rational, even when no other action would have been rational. This does not import any non-physical determinism. The mathematician is not forced to act as she does by considerations that exist within the space of reasons. She could write something different in a given line of the proof. That would, however, not be rational, unless she was deliberately composing an inadequate proof in order to see whether people would spot the mistake. Absent such unusual intentions, she can only act rationally by doing, or trying to do, precisely what she does in writing out a correct proof. (The qualification that she

might only try allows for the fact that someone who makes a mistake can still be acting rationally.)

The constraint to only one possible rational action is not one that commonly applies in its pure form. The constraint does not always apply to mathematicians when they wish to demonstrate results, because there may be alternative demonstrations. It does not apply when a mathematician is searching for a proof. It would then be perfectly rational to write down all sorts of things that would not in the end feature in a proof. In the natural sciences there may be many ways to establish results, and many experiments that are worth trying in the search for results. And in everyday life, a wide range of alternative rational actions is normally open to us.

Despite the contrast between the example of a mathematician writing out a proof and much of life, the example does make a point that is of general application. An action's rationality does not require a freedom to have done otherwise that is given at the level of reasons. That is, there is no need for there to be two or more alternatives that are acceptable by reference to reasons. It may seem obvious that a freedom to have done otherwise that is given at the level of reasons is not essential, so that an action can be rational even when it is the only rational action. But the point needs to be noted, because it follows that we do not need to worry that physical determinism might make it impossible for actions to be rational. We need not worry because rationality can exist even when there is no choice at the level of reasons. Physical determinism could not make such a lack of choice any worse. It would rule out other actions, but they would have been irrational anyway, and their being ruled out would hardly be a threat to the rationality of the actual action.

Although the existence of alternative possibilities might not matter for the rationality of a choice of action, it still seems that it ought to matter for the fact of choice. What could single out actions as chosen, apart from, at a minimum, our ability to see that the agent could have gone in a different direction or, if it we could not see that, the presence of some special factor that explained why we could not see that? I mention special factors in order not to require that whenever there is choice, there must be at least the appearance of alternative possibilities. Harry Frankfurt describes a disturbing case ("Alternate Possibilities and Moral

Responsibility", section 4). Jones chooses to do X, without any intervention, but Black would have intervened surreptitiously to change the workings of Jones's brain and body if Jones had been on the point of choosing to do Y instead. This case is disturbing because Jones does choose, even though we cannot see him as having been able to choose differently. But such cases are special and contrived. We can only see them as examples of choice against a background of more straightforward examples of choice where we can at least suppose that the agent could have chosen differently. The contrived nature of Frankfurt's example does not vitiate its philosophical significance, but it does limit that significance when we are concerned with choice, rather than with the moral responsibility that was Frankfurt's immediate concern.

No natural feature of the world looks promising as a mark of choice. Within the world as described naturalistically there is only what happens, and perhaps what might have happened if random events had turned out differently. A purely naturalistic description of the world, whether fully detailed or partial, makes it impossible for us to see that a putative chooser could have deliberately chosen differently, although she might have stumbled into something different. She could have done something different with a sense of deliberate choice because some earlier event had turned out differently as a result of randomness, but by the time of choosing, that event would have taken place and its outcome would have been fixed. There would still have been no controlled choice that was open at the time of choosing, nor could we see the putative chooser as having had options that were open to her. The different action would have been the only option in the revised circumstances.

Instead of looking at natural features of the world, we should seek other grounds for regarding certain conduct as chosen. But these grounds must not allow arbitrary attributions of choice. One attraction of linking choice to natural features of the world is that it would prevent arbitrary attributions. If something had to be found, we could not cheat and imagine that it was there. Can we do just as well even if we do not rely on natural features of the world?

We cannot do quite as well, but we can do something by recognizing our own sense of choice and then recognizing that others are like ourselves.

My proposed criterion of an action's being chosen is that it should follow from an episode that we see as being, to the putative chooser, subjectively like the episodes that occur when we make choices. We know what it is to make choices. We have just as clear an internal sense of choice as we do of action. We naturally take it that other human beings make choices in the same way that we do. Indeed, we need to do so. To refuse attributions of choice to someone would be to distance ourselves from him, to treat him as not human, because we would refuse to interpret his conduct in the way in which we standardly interpret the conduct of human beings. That sort of distancing is not ruled out by logic. We might do it occasionally, perhaps in extreme forms of cases of the sort that Peter Strawson cites as making an objective attitude rather than a participant attitude appropriate, cases of abnormality or of immaturity ("Freedom and Resentment", page 9). But we could not practise such distancing universally without forswearing the concepts and attitudes on which our lives as social beings depend. We would become like the objective anthropologist who is too remote to understand, and who "gets a sort of stony, distant look on his face" (Pirsig, *Lila*, page 33). Nor are we in a position even to consider practising such distancing universally, at least not without stepping well outside our normal habits of thought.

The view that other human beings make choices in the same way that we do is not just essential. It is also well-supported by the facts that we are physically very similar to one another, both internally and externally, and that we successfully relate to one another in a social world where it is taken for granted that people choose what to do. Indeed, the prevalence of individual choice is so fundamental a feature of our lives that we rarely notice it. In order to notice how all-pervasive choice is, we must contrast human beings, who have very diverse lifestyles even within a single town, and many opportunities to change their lifestyles, with other social animals, even chimpanzees, who have much more limited lives. The explanation of the contrast lies in our innate abilities and in our history. But that naturalistic explanation does not detract from the weight of evidence that other people make choices in the same way that we ourselves do. One could not live in a human society without continually making choices. Even someone whose life settles into a stable routine makes trivial choices, such as choices of what to eat or of whether to stay indoors when it is raining.

And settling into a routine will itself have involved choices.

There is no local inevitability about an individual's having a given way of life, no inevitability that springs from that individual's own nature and his immediate environment. Some individuals are predisposed to stable routines in general, and others are predisposed to instability in general, but that does not lead any given individual into a particular routine or into a particular sequence of changes. Influences that originate beyond the immediate environment may steer the individual into one routine, or one sequence of changes, rather than another. There may be global inevitability. If the whole Universe is deterministic, it will be inevitable that an individual should have a particular life, determined in all of its details. And all of the influences on a human being will have been transmitted to that human being or to his immediate environment through long causal chains. Exact reproduction merely of the last segment of each causal chain would lead to a physically identical life. But when we think about what an individual does, and about whether a human being should be seen as making choices, our vision is already narrowed to that human being and his immediate environment. It does not encompass the whole history of the Universe that lies on or within the light-cone that reaches back from him. All of that history would be implicated in determining the precise nature of the last segments of the causal chains that reached him.

This lack of local inevitability is related to an important role of the notion of choice in making the conduct of human beings comprehensible. We find ourselves in such diverse, changing and unexpected circumstances that it is implausible to think of ourselves as having tables inside our heads of all possible circumstances and of appropriate responses, prepared in advance, or even as having algorithms that could generate any required entries in hypothetical fixed tables on demand. We have to make up our lives as we go along. The scientific explanation of how we do that can be just as straightforward as the scientific explanation of how a computer can play a game of chess against any opponent that it meets and can improve with practice, even though when it was programmed no-one knew which opponents it would have to face, and even though it does not contain a table of all possible games or an equivalent algorithm. But the scientific explanation of the behaviour of a chess-playing computer need only refer

to the internal workings of the computer, clearly defined in terms of circuits and programs, and to the inputs that have been received, with those inputs all being in a very well-defined form, the moves that were made in games of chess. Human beings are much more complicated, their brains do not work in the crisp and well-defined ways that digital computers work, at least not unless we dig down far below the level of neurons and look at individual molecules, and the environmental influences are not at all precisely defined, again unless we dig down to the level of individual flashes of light, sounds and touches, or perhaps even deeper. The concept of choice allows us to simplify and to tell a comprehensible story that paints a clear picture of someone's life, whether over a day or over several years. We say that someone made certain choices, choices that we may well be able to relate to his character and to his previous choices, and that he acted accordingly. We make up for the lack of detail by using our intuitive sense of how human beings behave, and therefore of what makes an account of a day or a longer period a coherent story. This does not make it possible to predict anyone's conduct with great confidence, but it does allow us to make sense of conduct in retrospect. It is not surprising that historians who are not in the grip of fashionable theories recognize the importance of choices that are made by individuals (Roberts, "Postmodernism versus the Standpoint of Action", pages 256-260).

Here we find the first control over attributions of choice. It parallels the control over attributions of subject origination that was described in section 3.6. We should only attribute choice when the attribution plays a valuable role in giving an account of the conduct of the subject. We should not, for example, attribute choice to someone who is sleepwalking, or whose conduct springs from post-hypnotic suggestion. But even after such special cases are ruled out, the bounds of acceptability are wide. Choices that most of us would regard as crazy are still choices. The bounds of choice will be set wider than the bounds of rational choice, so long as we accept that a choice that is made for no reason, or one that is made contrary to what the available reasons would suggest, is still a choice. We can and should accept this. Someone can recognize that he makes a choice even when he has no reasons-based answer to the question, "Why did you do that?". We might not, however, accept the attribution to a creature of

choices for no reason or contrary to reason, when such supposed choices were the norm and there were few if any choices for reasons. A creature's choices that are not made for reasons may need to be in a minority for them to count as choices, because their status as choices may be parasitic on the status of similar episodes that are choices for reasons.

In section 3.4, I discussed the attribution of subject origination to computers, to robots and to aliens. The attribution of choice follows a similar pattern. There is not much point in attributing choice to our computers or to our robots, because we know very well how they work their way from inputs to outputs. There would probably be great difficulty in attributing choice to aliens, because their outer lives and their inner experience would probably differ greatly from our own. Here we see the second control over attributions of choice, a control that amounts to a re-statement of my proposed criterion of choice. Choice is only to be attributed if the attributor sees an episode that she can regard as being, to the putative chooser, subjectively like the making of a choice by herself is to her. We probably could not do this with aliens, because their inner experience would probably be too different from our inner experience. Likewise, there are limits to how far we can go with other animals, both because we cannot be confident of our grasp of whatever inner experience they might have, and because we have no idea what it would be like for our mental horizons to be as limited as their horizons are. We know what it is to be excited, or to be hungry, so we may have that much in common with other animals. But we cannot know what it would be like for feelings at that level to be the entire content of our mental lives, so that we would not, for example, have a sense of the difference between next week and next year. I claimed in section 4.2 that rationality was, in practice, rationality for us. Likewise, choice is choice for us. We attribute choice where other beings would not do so, and they would attribute it where we would not do so. This is an application to choice of Wittgenstein's remark, "We only say of a human being and what is like one that it thinks" (*Philosophical Investigations*, part 1, section 360, page 96).

To expand on the argument of the previous paragraph, similarity of inner experience matters directly in relation to attributions of choice, in a way in which it did not matter in section 3.3 when the task was to enter

into the heads of deliberators. The nature of a deliberation is largely given by the thoughts that constitute the deliberation. Those thoughts can often, although not always, be given independently of the nature of the inner experience of the thinker. A choice is different. It is a single momentary episode, the nature of which is given by reference to its phenomenology and to the fact that its effect is to narrow down a list of options to just one option. To see an episode as a choice is to see it as something, the chooser's experience of which is like our experience of choice. Only that perceived commonality of experience gives substantial content to the concept of choice. Thus similarity of inner experience is in general a precondition of attributions of choice, although it would be over-ambitious to claim that such similarity was logically necessary.

So there is for us a choice when there is an episode that precedes or occurs alongside an action, and it makes sense for us to see that episode as the making of a choice, both because seeing it like that plays a valuable role in giving an account of the subject's conduct, and because we can impute subjective similarity to our own episodes of choosing. We directly recognize such episodes in ourselves. Our immersion in society means that we directly recognize them in other people. We do not stop to look for evidence of choice, but take it for granted that other people choose in the same way that we do. We might recognize choices being made by non-human beings, although we cannot have any confidence that we could do so, both because we have not met non-human beings who are as sophisticated as ourselves, and because even the interpretation of the mental life of our closest relatives, the apes, is fraught with difficulty. We can get an idea of the level of difficulty by considering the meticulous analyses, and the many remaining uncertainties, that are set out by Barbara King in her book, *The Dynamic Dance: Nonvocal Communication in African Great Apes*.

In reflecting on the possibility of seeing others as making choices, it is important to heed the words, "it makes sense for us to see that episode as the making of a choice". I do not claim that a choice is an episode with a set of natural properties that make it a choice, properties that would be detectable by any observer of natural features of human beings and of the world who had the appropriate equipment. An episode is, to us, a choice

if we can sensibly see it as one. It and we between us do have natural properties that might well allow any observer who had both the appropriate equipment and an adequate ability to interpret states of our brains to say that we would regard it as a choice, but that would take us on to properties of ourselves. It would also still leave the putative choice as a choice for us, not a choice in itself. (The requirement for the observer to be able to interpret states of our brains takes us back to the discussion in section 3.7. Not just any observer of natural features would do. An observer whose own inner experience was too different from that of human beings would not be able to interpret states of human brains in the required way.)

Choice justified by reasons

If the concept of choice is to underpin the concept of rational action, it must be the concept of choice for reasons. A reason that plays the sort of role in the process of choosing an action that allows us to see the action as rational, in the sense of seeing that the choice was arrived at in an appropriate way, has to be a reason that the agent can articulate on demand. Some things are meant to be excluded by this formulation. If, for example, someone's brain is wired up in such a way that he will always reach for anything sweet, a state of his brain of which he is unaware, that fact about his brain should not count as a reason for his choosing to eat chocolate on a given occasion, although the fact will explain to others why he eats the chocolate. The state of his brain is not a reason in a way that would make a given act of eating chocolate an instance of his choosing to eat chocolate for reasons. A reason that the agent can articulate on demand, on the other hand, has a power of justification in the view of the agent, whether or not others would be sufficiently impressed by the justification to regard the action as rational. That is, the agent would consider statement of the reason to be an appropriate way to answer the question, "Why did you choose to do that?".

Some such reasons could be natural features of the world. The chocoholic might become aware of the state of his brain, and might then cite that state as the reason why he ate some chocolate. (The reason would become articulable when he actually became aware of the state of his brain.

It would not be articulable, in the sense that I intend, when he was merely capable of becoming aware of the state of his brain.) That state would then have justificatory power, to the same extent that "I did it because that is the sort of thing I like to do" has justificatory power. At least, it would have justificatory power to the same extent unless the inclination to eat sweet things led the agent to act against the actual or potential outcome of reflection on what he really wanted to do. An inclination that existed independently of reflection on what the agent wanted to do, but that inclined the agent to act in ways that were consistent with the conclusions of reflection that he had undertaken or that he would undertake if he bothered, would not be debarred from having justificatory power. This would be so even if the inclination was irresistible, unless the reflection really amounted or would amount to rationalization.

It would be unfortunate if we were to debar our pre-reflective inclinations, which are consequences of our natural features, from having justificatory power. If we limited justificatory power to preferences that were chosen from scratch, rather than admitting inclinations that existed in advance but that had passed muster, or that would pass muster, before the court of reflection, then the range of actions that we could justify would be implausibly narrow.

Whether or not reasons for a given action had their origins in natural features of the agent, their justificatory power could not be equated with natural features of the agent or of the world. An agent's perception of a reason as having justificatory power would correspond to natural features of himself, and those features would be visible to an observer of natural features who was able to monitor and interpret the states of the agent's brain. (As with attributions of choice, not all possible observers would have the necessary ability to interpret.) Likewise, the influence that the perception of justificatory power had on the agent's conduct would correspond to natural features, and would be visible to, and explicable by, any suitable observer. But the power itself would not be visible to all such observers, and it would not be the same thing as any set of natural features. Justification for actions amounts to their legitimation. This is something that a naturalistic description cannot do. A naturalistic description simply tells us how things are. It is devoid of the specifically human

presuppositions that lead us to see that certain actions are justified in certain circumstances.

A response to this line of argument would be that there is indeed no such thing as justification for actions, and that there are only delusive feelings of justification. One could take that line, but doing so would, like limiting ourselves to a detached psychological theory, leave us with accounts of our lives that would strike us as hopelessly inadequate. We do have a clear sense of justification for our actions, and individual justifications are given by our reasons for our choices of action. Our sense of justification can be unsettled in individual instances, when we discover that the reasons that we give for actions are not accepted by others as adequate justification. We may think again, and either amend our reasons or withdraw claims that given actions were justified. But very often that does not happen. The frequent acceptance by others that our reasons give adequate justification reinforces our sense that our reasons do indeed have justificatory power. The facts that any proffered justification might be rejected by others, that on reflection we might accept their rejection and that we cannot reliably predict the occasions of rejection, only show that we can make mistakes in believing that given reasons justify given actions.

CHAPTER 5

Knowledge

This chapter is concerned with epistemological questions. In section 5.1, I explore the source of those questions and set out the high price of avoiding them. We would have to keep propositions, as potential objects of propositional attitudes, out of the picture. That would in turn require an inhuman detachment from those around us. In section 5.2, I identify a connection between seeing ourselves as deliberating with mastery and the concept of knowledge. In section 5.3, I make a move from the inner to the outer that parallels the move that was made in chapter 4, by considering criteria for a belief to count as held rationally. The ability to respond to challenges is central here. After arguing for the importance of this ability, both as evidence that a belief has been acquired rationally, and as having an intrinsic relationship to our standards of rationality, I consider the possibility of rational computers. I then consider the sources of our standards of rationality. In section 5.4, I consider scepticism. I start by identifying one type of scepticism, the claim that we cannot have satisfactory justification for our beliefs, and the bearing of our concept of rationally acquired belief on that scepticism. I then argue that if the sceptic is to unnerve us, which is the most that he can do if we forswear certainty, he must use the purely structural form of his argument. I then argue that he fails because he is insufficiently liberal in the variety of structures of justifications that he will admit. Even circular arguments can provide justification, so long as the circles are large enough, despite the consequent potential for equally good justifications for contradictory conclusions.

5.1 Epistemological questions and their source

The traditional, and still unresolved, questions of epistemology cover a wide field. What are the conditions for a belief to count as knowledge? In particular, how does the traditional analysis of knowledge as justified true belief need to be modified? How, if at all, should the sceptic be answered, whether his scepticism be local, related to specific types of knowledge, or global, putting all or most of our claims to have knowledge at risk in one fell swoop? How may knowledge be classified, as a priori or a posteriori, or in some other way, and what falls within each category? In this section, I shall investigate the source of questions like these and whether we might avoid them altogether.

A necessary condition for questions that are related to knowledge to arise at all is the existence of things that might be known. Propositions must exist and must be potential objects of propositional attitudes, such as the attitude of belief. (I do not wish to pre-judge the sense of existence that is at stake here. The point is meant to be taken in a way that would be acceptable to those who were not Platonists about propositions. But I do need to take it that we have relationships, of belief and the like, to propositions. It might be possible to re-phrase the account that I shall give to accommodate, for example, a paratactic theory of indirect discourse, but it would certainly be awkward to do so.) To put the subject matter of epistemology in the broadest terms, it is some of our relationships to information that we consider as information. (I shall for convenience use the term "proposition" to mean a formulated piece of information, whether correct or incorrect. It may or may not be a complex of smaller pieces of information.) To understand where the questions of epistemology come from, and the conditions under which we might be able to block them at source, we need to understand which conceptions of the world and of ourselves must bring propositions into the picture in such a way that they are potentially objects of propositional attitudes. Being in the picture will mean being a potential topic of discourse. Mentioning, rather than merely using, propositions would therefore mean that they were in the picture. Being in the picture will not necessarily mean being seen as an object in the world, although anything that is so seen certainly will be in the picture.

If propositions are in the picture in such a way that they are potentially objects of propositional attitudes, that will open up a route to epistemological questions. If such a route is opened up, we should not then take the unprincipled option of avoiding the questions simply by refusing to reflect on our knowledge. My general approach has something in common with Ernest Sosa's distinction between animal knowledge, mere apt belief, and reflective knowledge, apt belief, the aptness of which the subject can defend against sceptical doubts (*A Virtue Epistemology*, page 24). If defence is a possibility, we should think about the implications of its being possible. Once reflection on our knowledge is an option, because we have brought propositions into the picture in an appropriate way, we have no principled choice but to take up the option. But if we did not have to bring propositions into the picture, or not in an appropriate way, we could decline to bring them in, or to do so in such a way, on the principled ground of an application of Occam's razor. We could then legitimately avoid the questions of epistemology.

I shall speak of propositions being exposed to propositional attitudes, as shorthand for their being exposed to being objects of propositional attitudes. Whether propositions need to be brought into the picture in ways that would expose them to propositional attitudes will depend on our conception of the world. We can start with the purely physical conception, in which we see only particles, forces and probability distributions. We can then go on to admit the full range of naturalistically defined entities and properties, apart from those that are defined within psychology. If we go no further than that, we have what I shall call the intention-free naturalistic conception. Using this conception, we do not attribute beliefs, desires and the like to human beings, so we do not take up the intentional stance toward them. It is to be distinguished from the intention-laden naturalistic conception, in which we do take up the intentional stance, but remain naturalistic by not attributing subject origination. These two forms of naturalistic conception of human beings would be available to far wider ranges of rational beings than ourselves, even if beings who did not have inner experience or outward behaviour that was like our own would have to rely on guesswork in order to identify an appropriate intentional stance to adopt toward human beings.

We use propositions in formulating a picture of the world under the intention-free naturalistic conception. We also find that we have to mention them. There is an important level of sophistication, essential to the practice of our sciences, at which we pick out regularities and at which we say that the same types of thing happen in different places. This involves mentioning propositions, but in a harmless way. We may pick out a mathematical relationship and refer to it, or we may comment that the same type of mathematical relationship occurs in different circumstances. In so doing, we mention a proposition, one that defines a variable by giving a specific equation, or we mention a specific type of proposition, such as linear equations, but we need not mention the proposition or the type in a way that would expose any proposition to propositional attitudes. "Here the equation is $y = 3x + 4z$", and, "All of the variables in this system are related by linear equations", are both propositions that can be objects of propositional attitudes if we bring them into the picture. But the proposition that is mentioned, $y = 3x + 4z$, and propositions of the type that is mentioned, linear equations, are not themselves exposed to propositional attitudes, because of the embedding within larger propositions. We can keep epistemological questions at bay, so long as we do not need to mention the larger propositions as wholes. We can avoid doing so if we can state the content of our science with all propositions that are mentioned, either directly or implicitly through mention of their types, being buried within propositions that are merely used.

Holding this line between, on the one hand, use and mention that is out of the reach of propositional attitudes, and on the other hand, mention that is within the reach of propositional attitudes, does require the drawing of a clear boundary between the observer, who uses propositions, and the world. If the observer were regarded, by himself or by us, as being in the world in his capacity of observer, then the propositions that he used would have to enter the picture in ways that would expose them to propositional attitudes. He would be seen as observing that p, or that not-q, in a way that made room for explicit assent. If he were not seen in that way, with room for explicit assent, then he would not be seen as an observer, but merely as an object in the world that changed its internal states in ways that reflected other changes in the world. If, for example, a red light came

on, his brain would change in certain ways. It would be a step beyond that, bringing propositions into the picture in ways that would expose them to propositional attitudes, to say that he observed that a red light came on. We can, however, draw the boundary between the observer and the world, with just one reservation. The reservation is that there are delicate, and still unresolved, questions about the role of the observer in connection with quantum mechanics.

Leaving that reservation to one side, all that we would need would be a conception that allowed us to view the world as an object of study, from which the observer as observer was excluded. Propositions that were entertained by the observer would not then be in the picture, even though the observer's brain cells would be in the picture. I shall refer to the objective conception of observation to capture the idea of the world's being an object for the observer, who is not, as observer, in the picture. This is a matter of the relationship between the observer and the world, rather than a matter of how the contents of the world are seen. The intention-free naturalistic conception, the intention-laden naturalistic conception and the originator conception, on the other hand, are conceptions that allow us to see the contents of the world in particular lights.

Use of the intention-free naturalistic conception of the world would allow us to avoid bringing propositions into the picture in a way that would expose them to propositional attitudes, so long as the process of observation could be brought under the objective conception. (If we actually thought about the process under that conception, rather than merely regarding the process as apt to be brought under that conception, then the questions of epistemology would be brought down on us. This would happen because we would think about the observer and his relationship to the information that he gathered. The words, "could be brought under the objective conception", must therefore be read as, "could be brought under the objective conception by some hypothetical super-observer whose thoughts we did not reproduce or contemplate within our own minds".) Use of the intention-laden naturalistic conception would not allow us to avoid bringing propositions into the picture as potential objects of propositional attitudes, because its use would involve attributing beliefs to people. We could then ask whether they knew the propositions that they

believed, and what types of knowledge they could have. Use of the originator conception for the purpose of seeing deliberation with mastery would present the same danger. We would see people as standing back and considering pieces of evidence and possible methods of argument before deciding how to proceed to conclusions. That would bring propositions into the picture as objects of propositional attitudes. We would then have to be ready to tackle epistemological questions. Whether we can avoid epistemological questions therefore hinges on the extent to which we can make do with the intention-free naturalistic conception. We can get an idea of how difficult it would be to avoid including propositions in the picture as potential objects of propositional attitudes by considering a few examples where the importance of propositions is, to us, manifest. My chosen examples are laws of nature, the data, programs and output of computers, and human thought.

Laws of nature

A law of nature is, at the very least, a statement of an observed and predicted regularity. It may well be more than that, although philosophical opinions differ as to what, if anything, the additional properties of laws might be. A purely physical description of the world, giving the locations and movements of particles or other information at that level, might not appear to be able to disclose law-like regularities. It might seem that the regularities would be present, but unnoticed. But that would be too narrow a view. Propositions that recorded the regularities could be formulated and included in the description, so that they would be mentioned, but as noted above in connection with mathematical equations, they would not need to be mentioned in such a way as to expose them to propositional attitudes. (The advancement of science, as opposed to the mere statement of current knowledge, would require us to expose propositions to propositional attitudes, because we would have to debate which propositions were true, and by implication, which ones we knew or believed. At least, that would be so unless we could do all of our sorting of true propositions from false ones at some sub-personal, non-self-conscious, level. And we would certainly need to expose propositions to propositional attitudes if we wanted to say things like, “All of the results that have been obtained in John’s laboratory are correct”.)

There is no reason to think that we would need to expose propositions to propositional attitudes in order to accommodate additional properties of laws of nature. The most likely additional properties are some form of necessity, and membership of a set of propositions with some specified characteristic. Possible characteristics of sets of propositions include that of being well-structured by relationships of implication, and that of being stable in the face of the outcomes of any circumstances that could transpire given the truth of all members of the set (Lange, *Natural Laws in Scientific Practice*, chapters 2 and 3). One would have to mention propositions, thereby bringing them into the picture, and not merely use them, in order to notice necessity or membership of a set of propositions with some given characteristic. Furthermore, one would have to notice necessity in order to enquire into its nature and sources, and one would have to notice membership of a set of propositions with a given characteristic in order to consider the significance of that membership for the status of a proposition. But this does not mean that considering the possibility of holding propositional attitudes toward the propositions in question would help to explain either necessity, or the fact that propositions were members of a set with a given characteristic. This lack of explanatory usefulness would make adequate an approach that kept the mentioned propositions at a distance from propositional attitudes. We could do all that we needed to do in order to state our current knowledge merely by using propositions such as, "Necessarily, an object that is released in a gravitational field is accelerated", or, "The proposition that when hydrogen is mixed with oxygen under appropriate conditions, water is formed, is a member of our set of natural laws". Such propositions as wholes would merely be used. They would merely play roles in forming our expectations about the world, and in guiding our conduct. The propositions that were embedded within them, and mentioned, would be kept at a distance from the surface, so that they were not potentially objects of propositional attitudes. Likewise, an analysis of laws of nature that did not attempt a reduction to non-nomic properties, such as the analysis that is proposed by John Carroll in "Nailed to Hume's Cross?", would not require us to mention propositions in such a way as to expose them to propositional attitudes, so long as we only wanted to state some of our current knowledge by identifying certain propositions as laws of nature.

We must consider an objection to this approach. The proposal appears to be to use some proposition, p, and also to use the proposition, Necessarily(p). How could we recognize that the same proposition, p, was used in the former case and mentioned in the latter, without mentioning p in isolation? The answer is that this task of recognition does not need to be performed. Once we recognize that we have a law, and move up to saying "Necessarily(p)", we stop saying "p". That is, we do not need to make the connection between the two propositions, because the new one supplants the old one.

Our computers

I shall now turn to the data, the programs and the output of computers. When we look at data, programs and output, we know what the data represent, what the programs do and what the output tells us. We might, for example, see the data as figures for births in different regions in a given year, and a program as a way of estimating the need for schools in those regions five, ten and fifteen years hence. Propositions would be seen as expressed over and over again, in ways that exposed them to propositional attitudes. A datum might be seen as expressing the proposition that 100,000 children were born in region B last year. A program might be seen as expressing the proposition that if the numbers of births in regions B, C, D and E in a given year were w, x, y and z, then the number of children starting school in region C five years later would be some specific function f(w, x, y, z). Some output might be seen as expressing the proposition that we would need six more schools in region D ten years from now. The meanings of data, of programs and of output are obvious to us because we collect data, and write programs, in order to get the output that we want. All of the data, the programs and the output are invested with meaning, and can be explained. Sometimes the explanation of a program is merely that someone was having a bit of fun practising his programming skills, but that is still enough to make sense of the program. We can arrange for data to be collected automatically, and for them to be processed using programs that a computer has devised or modified without reference to human beings, but even then, we can understand what is going on by tracing the story back to human beings who wanted to do something.

Those human beings could in turn work their way forward through the story and interpret what was going on, just as if they, rather than the computer, had written the programs and collected the data. They would then see the data, the programs and the output as expressing propositions. They would also see those propositions as exposed to propositional attitudes.

In the light of this, how could we keep epistemological questions at bay? The origins in human intentions of programs, even those that are written by computers, and of the collection of data, mean that we could only do so if we could also avoid bringing propositions into the picture as potential objects of propositional attitudes when we considered human beings under the naturalistic conception. The link between human intentions and computer programs is close enough to make that a precondition of success. I shall turn to human thought shortly. If we could prevent propositional attitudes from getting a grip in connection with human thought, that should suffice to dismiss the claim that the "thoughts" of computers when they themselves wrote programs, collected data and "interpreted" output would require us to bring propositions into the picture as potential objects of propositional attitudes. Meanwhile, considering computers in isolation, it might seem to be straightforward to keep propositions out of the picture entirely. A computer is a physical object. Electrical charges enter it, move around within it and emerge from it. The processes of input and output are also physical, involving the movement of keys and the distribution of electrical charges on a screen. We could describe everything in purely physical terms, without loss.

To be precise, we would not lose information by falling back on a purely physical description, one that did not mention the propositions that we in fact used when talking about the data, the programs and the output. A purely physical description would capture all of the complexity that a description that mentioned those propositions would capture. (In fact it would capture more complexity, because we are normally highly selective in what we describe, overlooking most of the physical detail in the world.) Every detail would be captured in the sense that if the set of propositions that could correctly be asserted were different, something in the physical description of the world, either inside or outside the computer, would also

be different. But we would lose intelligibility. The most natural and informative way to describe what goes on in and around a computer is to state the propositions that are expressed by the data and the output, and to state what is accomplished by the execution of its programs. What is accomplished can be expressed by stating propositions that set out the functions from inputs to outputs, and adding that the execution of programs involves inserting the inputs and calculating the corresponding outputs. So must propositions come into the picture in ways that would expose them to propositional attitudes?

When we fall into describing the world, excluding human beings and perhaps some higher animals, as if the entities in the world had propositions in mind or were deliberately carrying out certain tasks in order to achieve goals that they had formulated, we must remind ourselves not to anthropomorphize. But if we describe human beings without mentioning their beliefs and intentions, as propositions in their heads, then we lay ourselves open to Sidney Morgenbesser's reported characterization of the views of the behaviourist B. F. Skinner: "You think we shouldn't anthropomorphize people". Computers seem to fall somewhere in between, but closer to human beings than to the rest of the world because we design, build and program computers, and feed them with data, in order to accomplish tasks that we have chosen. We can even see computers as extensions of ourselves that are active participants in our mental processes, rather then their being mere external tools (Clark, "Reasons, Robots and the Extended Mind", section 4). Such a view might tempt us to anthropomorphize computers, although that temptation could be resisted because if a computer is seen as an extension of oneself, it is not granted the status of an independent being. (Resistance becomes harder if a computer is seen as an extension of a group of human beings who are collaborating on some project, because it is not then seen as an extension of any one human being. It can then easily be seen as an additional member of the group.) Regardless of whether we should see computers as extensions of ourselves, a severe loss of intelligibility would result from not mentioning propositions when explaining what computers did. We would suffer this loss even if the computers in question wrote their own programs and collected data themselves, having been created by us with

the ability to perform those tasks or to learn how to perform them. This amounts to a serious challenge to the idea that we could have an adequate conception of the world that did not mention propositions in ways that exposed them to propositional attitudes. It is a serious challenge because the propositions would certainly be exposed to propositional attitudes if they were entertained by the people who had set the computers their tasks.

The challenge can be met, but when we consider the terms on which it can be met, we will see the high price of keeping epistemology and its questions at bay. The price will turn out to be the maintenance of a rigidly detached stance, from which we do not see the programmers and users of computers as entertaining propositions. It is hard to imagine detaching ourselves from the human world to this extent, but it is just about possible to imagine doing this thing that we almost certainly cannot do. The reasons for looking at computers, rather than moving directly to human beings, are that it is easier to imagine our achieving the necessary detachment if we focus only on our computers, and that looking at the relationships between computers and their human programmers and users highlights the significance of the specifically human element. Another way to help us to imagine achieving the necessary detachment is to consider encounters with aliens and with their computers, in which detachment would be forced on us. Identifying the reason why detachment would be available to us in relation to them will help to bring out the cost of detachment in relation to human beings.

Aliens and their computers

Suppose that we came across things that some aliens used in the ways that we use computers, and suppose that we saw those things in action, but that the way of life and patterns of thought of the aliens differed greatly from our own, as they probably would. We might well suspect that the things we had found were analogous to our own computers, but we would be at a loss to impute content to the processes that went on inside them. We could observe and describe the physical components and the electrical impulses, or some other mechanism of state transformation if different technology were used, but the differences in way of life and patterns of thought of the aliens would mean that we could not impute propositional

content, or even assume that it was appropriate to characterize the processes in terms of propositions at all. We could give a global measure of the information content. We could probably go further and record certain regularities of input, process and output, but with no idea of what, if any, significance the aliens would attach to the things that were picked out by the terms that we used in stating those regularities. Despite this lack of grasp of what was going on, and even despite our awareness of the lack, we could give a full "natural" history of the objects that we had found by describing the components and impulses. This is something that we would consider to be perfectly adequate if we were studying animals that were much less sophisticated than ourselves, rather than studying artefacts.

Now suppose that we met the aliens who had built the computers. We could have the same attitude toward them. We could give a complete natural history of them, but we could not impute specific thoughts or intentions to them, or even assume that any such characterization would be appropriate, because we would not understand them at all. To recapitulate the discussion in section 3.7, philosophers have shown us how to make a start when parachuted into a strange society with a completely unknown language. It might be thought that those methods could be used here, but they could not. Use of the principles of charity and of humanity relies on an assumption that those who are to be interpreted have a fair amount in common with ourselves. We would have no right to use those principles when we were dealing with radically different aliens. We could try out various intentional descriptions of the aliens, and we might be fortunate in finding one that worked, in that it allowed us to systematize and to predict their behaviour. If we did so, we could reasonably construct a detached psychological theory of the aliens, in the sense that was given in section 3.7. But there is no reason to expect that we would be so lucky.

If we had a natural history of the aliens, and a "natural" history of their computers, histories that did not impute propositional content, would anything be missing? We would have a suspicion that something would be missing, but we would be at a complete loss to say what it was. Having reached such an extreme point, of having no conception of what might be missing, we would not be entitled to assert that nothing would be missing, but we could stop worrying about whatever it was that might be missing.

At most, we might be able to say that it was something that would, if it were known, appear to us as something like intentional understanding. What is more, we could stop worrying ad hoc, on the facts of this instance alone. We would not have to forswear ad hoc attitudes and take up a general ontological anti-realism, a view that since the putative missing thing would be outside our framework, the question of its existence did not even arise. (Such a general position would arguably be inappropriate anyway, because we might be able to see the missing thing as something that lay at an impossibly remote point on a scale, the near end of which was within our framework. This would be the scale of types of intentional understanding.) We can get a sense of the level of ignorance that we would have by reflecting that if the computers were made mobile and physically active, that is, if they were robots, we might not be able to tell whether an entity was an alien or a robot that had been built by the aliens. It would certainly be rash to decide on the basis of whether the important element in the entity was carbon or silicon, or whether the entity moved around on legs or on wheels, or whether entities of its type reproduced with one mother per child. The facts about life on Earth that might support the use of such clues would be too contingent to allow the clues to give us trustworthy answers on different planets. We could not confidently say which type of entity was the natural one, and which was the robotic one, unless we found the factory where the aliens made their robots, or worked out that entities of one type were too simple to have built entities of the other type, or even to have built entities that might in due course have given rise to entities of that other type.

Aliens might not be so different from ourselves. We might be able to comprehend them, at least to a limited extent. But the point that matters here is not whether there are in fact any wholly and permanently incomprehensible aliens. What matters is that the reason why we are entitled to say that nothing worth worrying about would be missing from our natural history of incomprehensible aliens, a natural history that did not impute propositional content, is that we could not grasp how they thought. We would be confined to the intention-free naturalistic conception. Thus we would not see the aliens as thinking in propositional terms, a way of seeing them that would open the way to our seeing them

as mentioning propositions and as taking up propositional attitudes. We would indeed have no conception of any propositional attitudes that they might have. Turning to our descriptions of them, they would be like naturalistic descriptions of the world generally. We would use propositions. We would also mention propositions as a way of capturing generalizations, but we could embed those mentioned propositions in propositions that we merely used, keeping the mentioned propositions safe from being potential objects of our own propositional attitudes. Epistemological questions could therefore be avoided without indulging in an unprincipled refusal to philosophize, so long as we could simply set down our descriptions, and did not have to debate which propositions stated truths about the aliens and which ones stated falsehoods. (As with laws of nature, debate at some sub-personal, non-self-conscious, level could be conducted without exposing propositions to propositional attitudes.)

The perils of detachment

We can now consider the option of keeping epistemology at bay even though we describe human beings rather than aliens, and the cost of keeping it at bay. I have already moved beyond computers into the realm of natural, but thinking, beings. We now enter the third area that I promised to tackle, human thought. We need to consider whether we could describe human beings in ways that we found adequate, without bringing propositions into the picture in a way that would expose them to propositional attitudes. If we could do that, we might keep epistemology at bay. We were safe with incomprehensible aliens, because we could not even characterize their thought in propositional terms. Human beings are different, because we do characterize their thought in propositional terms. We also see them as taking up propositional attitudes.

We might appear to have a choice, between not characterizing their thought in propositional terms, and so characterizing it but refusing to see them as taking up propositional attitudes. But the latter option would not be available. When we consider human beings, both their awareness of their own processes of thought and their self-consciousness must be taken as given. We know that once human beings use propositions, they can mention them, and that once they mention them, they can take up

propositional attitudes toward them. It would not even be necessary for someone who was seen as using a proposition to be seen as consciously taking up a propositional attitude toward it himself. We could do so on his behalf, imagining that we believed the same proposition and asking whether we would know it. That exercise would allow us to imagine that the person who was being studied consciously believed the proposition. The similarities between human beings would ensure that such an exercise was possible. So as soon as we see human thought in propositional terms, we are on the road to epistemology. (We cannot avail ourselves of the option of embedding all mentioned propositions in propositions that are only used, thereby putting the mentioned propositions out of the reach of propositional attitudes, because we must see the used propositions as open to being mentioned, and therefore as exposed to propositional attitudes.) If epistemological questions are to be avoided, we must not see people as using propositions that we understand. But that would imply an unnatural level of detachment from people. We would be as remote from them as we would be from incomprehensible aliens. Could we manage that level of detachment? And if we could, would we have an adequate view of human beings?

Two preliminary points should be noted. The first point is that the requirement would be to adopt the intention-free naturalistic conception. (We could not use an intention-laden conception but regard the propositions that people used as incomprehensible, because we know that we can understand the propositions that people use.) Use of the intention-free naturalistic conception is a matter of how the contents of the world are seen, so it should not be confused with the objective conception of observation, even though if we want to keep epistemology at bay, we must also regard observation as apt to be brought under the objective conception. Indeed, the observer of humanity who wishes to avoid epistemological questions must go further and describe herself, as well as other people, using the intention-free naturalistic conception, if she describes herself at all. The second point is that because the objects of study would be, like the observer, human beings, any need to see propositions as potential objects of propositional attitudes that were held by someone who was being studied would bring the questions of

epistemology down on the observer. If we can ask whether any human being knows a proposition that he believes, then questions as to the possibility of, and the conditions for, knowledge become pressing for all of us, in relation to all of our putative knowledge. This follows from the fact that we have so many features of our mental lives in common.

The answer to the first question, that of whether we could manage the level of detachment that would be involved in not characterizing human thought in propositional terms, is very likely to be that we could not. We could try to achieve such detachment, and perhaps we could manage it with people we had only heard about and had never met, nor ever expected to meet. But even then, it would be difficult. It is difficult to remain detached even when reading a biography of someone who lived hundreds of years ago. That is partly a consequence of the biographer's art. A good biographer will make figures from the past come alive for us. But even if we read a dull encyclopedia article about a historical figure, an article that merely recites a few facts, we see the person as one of us, so that we relate to him as a fellow human being. We do not achieve total detachment. Peter Strawson pointed out how difficult it would be to maintain an objective attitude toward people around us ("Freedom and Resentment", pages 9-10). I am discussing an attitude of much greater detachment than Strawson had in mind, because his objective attitude would still allow us to see the trains of thought of the people concerned in propositional terms. So complete detachment is probably beyond us. We do not, however, need to settle the question in order to pursue the argument. We can move on to the second question. Assuming for the sake of argument that complete detachment were possible, would we have an adequate view of human beings if we did not even impute propositional content to their cerebral activity?

The view of human beings that was so given would strike an observer as adequate, if and only if complete detachment into which the observer was locked were achieved. Given that an observer was locked into such detachment, she would be like we would be when studying incomprehensible aliens, in that she would have and could have no idea of what was missing. A description that imputed no propositional content would then be as adequate as any description could be, and she would be

unaware of its limitations in any more than the vaguest way. In fact, the hopeless inadequacy of the postulated description screams out at us. But that is because we are human beings who can see that other human beings are like ourselves. We have not achieved the complete detachment that would be required, so we can see that achieving it would entail a catastrophic loss of understanding. We can see that, because we can see in some detail what would be lost. We would lose the human point of view. A catalogue of the regularities in changes in the states of people's brains, correlated with regularities in their outward behaviour, which did not characterize their thought in propositional terms, would not allow any appreciation of what it was like to lead a human life.

Detachment of the type that is demanded here means seeing beings as mere physical objects in the world, and not as participants in our rational way of life. This explains why detachment is easy in relation to parts of the world that are devoid of human beings and of the kinds of animal that we are tempted to anthropomorphize. We have no temptation to imagine rocks or plants as being like us, so we can do geology and botany in a detached way, as pure natural history, without feeling that anything is missing. Not much changes if we move up the scale to insects. If we move up to more sophisticated animals, particularly those that have cultural significance for us, such as dogs and bears, the temptation to anthropomorphize grows. If we move up to human beings, detachment becomes practically untenable. It may not be absolutely untenable. Those who are committed to all three of methodological, psychological and logical behaviourism, and those who see human beings merely as huge buzzing systems of atoms, might harden their hearts and maintain their positions. But that would not be an easy thing to do, and the attitude would not endure once the hard-hearted left their studies and conversed with their friends.

Several philosophical views bear on the question of whether a detached view of human beings, in my very strong sense, would be adequate. Derek Parfit considers reductionist views of people, and ultimately accepts them (*Reasons and Persons*, part 3, with the conclusion on page 280). Such reductionist views would facilitate taking a detached view. Quassim Cassam argues that we cannot give an adequate account of

first-person thoughts without including persons in our description of the world, so that a reductionist description would inevitably be an incomplete description of the world (*Self and World*, chapter 5, section 3). We would also, in Cassam's view, lose the unity of a life (ibid., chapter 5, section 4). Lynne Rudder Baker argues for the vital role of a strongly first-personal perspective (*Persons and Bodies*).

My response to Cassam and to Baker is a shift of focus, rather than any disagreement. The content of first-person thoughts would not feature in a detached view that saw human beings purely as natural mechanisms, and that did not see their thoughts in propositional terms. Only physical states of human brains would feature. The incompleteness that concerns Cassam and Baker would be part of what we were trying to achieve, the elimination of propositions from our view of the world. If we were to achieve that, our view of the world would be grossly impoverished, but it would not be incomplete in the sense of missing out something that could have been included in a detached view of the type that I am discussing. Likewise, we would lose any sense of the unity of a human life as something that made sense rather than its being a chapter of accidents, and any sense of responsibility for past and planned actions or of answerability to ethical standards, but these senses of unity, of responsibility and of answerability are things that one would never expect to see if one took a detached view. One would only expect to see the creation, continuation, change and destruction of objects that were characterized in intention-free naturalistic terms. We can also note that if Cassam and Baker are right, and if in addition rational beings are sufficiently diverse that the contents of some thoughts of some beings could not be grasped by us, either while they were first-person thoughts in some sense, or while they played some equally vital role in giving complete accounts, then all of our descriptions of portions of the Universe that were large enough to include such diverse rational beings would be incomplete.

The inevitability of epistemology

To sum up the argument so far, the questions of epistemology can be dismissed if we can do all that we wish to do without mentioning propositions in ways that expose them to propositional attitudes. But the

price of avoiding epistemology would be a high one. It would be a wholly inadequate conception of human beings. It is not merely that an inhuman detachment would be imposed on us. We would also be unable to appreciate our rational way of life. When we advanced our knowledge by debating which propositions were true, and by implication debating which propositions we knew or believed, we would have to refuse to take note of what we were doing in conducting such debates. We would also be unable to see people as deliberating by reference to evidence that they considered, because that would involve seeing them as stating the evidence in the form of propositions. If we could not see people as considering evidence, then we would not be able to appraise actions as rational, because any such appraisal would require us to see the subject as having applied appropriate decision-making methods to considerations of which he was aware. Those considerations would have to be seen as being in propositional form in the subject's mind, in order for us to consider whether he had made appropriate use of them.

The inevitability of epistemology does not mean that we must worry about epistemology all of the time. Most of the time, we ignore the subject. We gather data, test our hypotheses and reason our way to conclusions without worrying about the philosophical legitimacy of our procedures, as distinct from their practical efficacy. If our procedures are known to yield only approximate results, or if we know or suspect that they are liable to error, we allow for that. We recognize the limits of what we do, but we do not stop to ask whether the results that we obtain qualify as knowledge, or whether the propositions that express the results are analytic or synthetic, in the way that epistemologists ask those questions. We may notice things that would also be relevant to the nature of our knowledge, for example as a priori or as a posteriori. We may, for example, ask whether our results follow from axioms of logic and mathematics, or whether we have made essential use of empirical data. But we make such enquiries in order to establish the degree of confidence that we may have in our results, and in order to work out what sorts of checks for errors we should make, rather than making the enquiries in order to lay philosophical worries to rest.

The fact that we ordinarily allow ourselves such a philosophically easy time does not mean that we should do so. While it would be excessive

to think about epistemological questions all of the time, there are good reasons to pose questions of a broadly epistemological nature every now and then. They prompt us to ask whether our procedures are really as good as we think they are. Do we systematically overlook certain data? Do our habits of thought lead us to ignore some possible interpretations of the data? Do the conclusions that we proudly announce have substantial content, or do we say very little because there are many free parameters in our theories that can be given values in order to make the theories fit the data? And so on. Such questions belong to scientific method, a discipline that is adjacent to epistemology and that shades into it, and they are undoubtedly worth asking. On the other hand, we would never get anything done if we worried about such things all of the time. And the questions cannot sensibly be posed in relation to the acquisition of everyday beliefs, about how our nearest and dearest are feeling or about the time of the next train, because the procedures by which we acquire such beliefs, and the contents of the beliefs, are either too ill-defined for the questions to be answered, or too straightforward for the questions to be serious.

5.2 Knowledge and deliberation with mastery

If we see ourselves as deliberating with mastery when we adopt some of our beliefs, that ensures that we refer to propositions when characterizing the mental processes of human beings, so that epistemological questions are unavoidable. Any characterization of our deliberations that saw us as considering reasons would have that consequence. But seeing ourselves as deliberating with mastery has a further effect. It has a particular influence on our concept of knowledge.

The traditional analysis of knowledge as justified true belief has been known to be inadequate since the publication of Edmund Gettier's paper, "Is Justified True Belief Knowledge?", in 1963. Other proposed analyses have mostly adopted one of two broad approaches. The first approach has been to require that justification be appropriately linked to truth. This is the approach of the prohibition on false lemmas, an approach that has recently been argued to be relevant even to some Gettier cases that had

been thought not to rely on false lemmas (Levin, "Gettier Cases without False Lemmas?"). The second approach has been to bypass justification in favour of some purely causal account, or to relocate it as justification for our use of the methods by which beliefs are formed, rather than as justification for specific beliefs. A leading strand within this second broad approach has formed around Robert Nozick's idea of tracking, the application of which also rules out some claims to knowledge that rely on false lemmas (Nozick, *Philosophical Explanations*, chapter 3, section 1).

Seeing ourselves as deliberating with mastery matters here because on some, but not all, occasions, we see ourselves as freely considering the weights to attach to pieces of evidence and the possible methods of argument from evidence to conclusions, before we work out which beliefs to adopt. If we ignored this view of the process, and always saw the acquisition of knowledge merely as the mechanical execution of programs that would produce correlations between states of brains and states of the world, or that would produce appropriate dispositions in brains to respond in certain ways to requests for information, that would not be satisfactory. We would not have a picture of our acquisition of knowledge that measured up to our self-conception, because we would not see a stage of free contemplation of the evidence and of how it might be used. We would see ourselves as no more than followers of programs, albeit sophisticated programs that generated reasoned discourses.

We should, of course, require the existence of correlations or of appropriate dispositions as a necessary condition of knowledge. If their material existence were required, that would amount to requiring known propositions to be true. If their counterfactual existence were required as well, that would amount to imposing a tracking condition on the processes by which knowledge was acquired. But it would not be satisfactory to see our acquisition of knowledge merely, and on all occasions, as the execution of programs that mechanically produced correlations or appropriate dispositions. As I shall argue below, we take some of our sophisticated knowledge to be obtained in ways that we see as going beyond the mechanical execution of programs.

Correlations and dispositions can be engendered in brains by means that do not even involve reasoned discourse. Any of our beliefs could be

engendered by subliminal propaganda or by drugs, but beliefs that were engendered in those ways would at least initially be regarded with suspicion by anyone who was aware of their origins, even if the processes by which they were engendered were considered to be reliable inculcators of true beliefs. Beliefs that were engendered in such ways could qualify as knowledge, so long as the processes were indeed reliable. Learning by rote, which comes closer to such ways than it does to the reasoned discourse that we expect in relation to sophisticated beliefs, does engender knowledge. I do not propose to incorporate a new condition in the definition of knowledge, a condition that a belief should be seen as adopted through a process that involved mastery, or even reasoned discourse without mastery. But there is one way in which we should see our concept of knowledge as influenced by the idea of deliberation with mastery.

The influence is that while large amounts of knowledge can be seen as engendered by mechanical processes (including reasoned discourses generated by programs), the concept of knowledge can be a concept of something rather more than true belief that has been acquired with due care. If we saw all of our beliefs as engendered mechanically, we would be limited to seeing anything extra, over and above reliably true belief, in the relationships between our beliefs. We could not see anything extra in the nature of the mechanism. The most that we could get from that would be something close to a guarantee of the correctness of the beliefs that were acquired by using the mechanism.

The relationships between our beliefs amount to more than consistency. The truth of some beliefs explains the truth of other beliefs, and there are relationships of logical implication between some beliefs. But if we see some beliefs as adopted through deliberation with mastery, that allows us to see something else, beyond the relationships between beliefs. It allows us to see the beliefs as adopted in ways that make use of a special capacity for free thought, involving a type of reflection that is beyond mechanical computation. If we recognize that we see at least some beliefs as adopted in that way, we can better understand why we regard knowledge as something special, as having intrinsic value beyond the mere true belief that Plato discusses (*Theaetetus*, 187-210), and why we esteem human beings for possessing knowledge and for increasing our stock of it.

The influence on our concept of knowledge is particularly strong in connection with the acquisition of academic knowledge that is new, not just to the acquirer but to humanity. If someone has learnt facts that are already well-established, it does not matter much if we do not see him as having adopted his new beliefs through deliberation with mastery. It does not even matter if they were engendered by rote learning, a process that patently does not involve reflection on the specific beliefs that are acquired. Likewise, the routine acquisition of data directly from the natural world, rather than from human teachers, can be seen as proceeding mechanically. But when we are at the frontiers of human knowledge, we cannot know whether any mechanical process would reliably guide us in the next steps, except in a few special cases, for example, when we have devised a computer program that generates new mathematical theorems, and we have confidence in the results because we understand how the program works. Given that there is in general no mechanical process, specified in advance, on which we can rely, we naturally expect to use reasoned discourse of a type that will allow us to break free of any particular inherited mechanism. We expect to be able to stand back and consider all of the available information, how to weight the evidence and which method of argument to use, in a way that leaves open all of the options that are in general available to us. We know that we cannot really do this, because we appreciate the causal structure of the physical world and the dependence of our mental lives on our physical brains. Despite these facts, we still only consider ourselves to be acting ideally in the enlargement of human knowledge if we do see ourselves as standing back in a way that leaves all of the options open, that is, if we see ourselves as adopting our new beliefs following deliberation with mastery.

In such contexts, we attach high value to new beliefs that we see as acquired in that way. Seeing deliberation with mastery helps us to see creativity. If the beliefs then stand up to tests of their correctness, we consider them to be particularly admirable examples of new knowledge. We would not be so impressed even if we saw a mechanical process of the utmost sophistication that took us through all of the possible sets of weights and all of the known methods, and that made selections with the greatest sensitivity to clues in the evidence as to what it would be best to do.

We even doubt the capacity of such mechanisms, because when we are on the frontiers of knowledge, we are aware of the risk of not knowing which clues are really significant. We can in fact do no better than use such a sophisticated mechanism, and furthermore a mechanism that is, at the highest level, not even chosen by us because the all-encompassing programs that capture the entire contents of sets of rational antecedents are simply foisted on us, but that is not how we want to see things. Our desire not to see things in that light is encouraged by our awareness that only some mathematical functions, the Turing-computable functions, can be computed using methods that we have already defined precisely. We have the hope, perhaps groundless, that other functions may be computable in ways that are as yet undefined. By analogy, we are aware that some leap of imagination that we cannot manage to see as emerging from a mechanism may very well be necessary in order to advance our collective knowledge, even though any actual leap would have to be made by physical entities that were embedded in the causal network.

5.3 Rationally held beliefs

At the start of chapter 4, I turned from our view of the nature of processes of deliberation to our appraisal of the content of the results and the content of the processes. I shall repeat that move now, leaving behind the nature of the process of adoption of beliefs and turning to the question of whether we should regard someone as holding a belief rationally. I shall take it that a belief is held rationally if it has been acquired rationally, and if any challenges to it that have so far been mounted have been seen off in appropriate ways, using counter-arguments that are generally accepted as rational rather than relying on counter-arguments that do not meet that standard or on a refusal to listen. What is at issue is the relationship between the beliefs that are held and the evidence that is available to the believer. It does not matter for this purpose if the process is seen as mechanical. But three other requirements must be met. There must be a process that is of sufficient complexity to support an interpretation that ensures that the pieces of evidence, the method of argument and the

conclusion are all represented. The interpretation must fit with the external facts about how the subject speaks and acts more generally. And application of the interpretation to the process must reveal the subject's reasoning as it is seen by the subject. (It must not reveal less than that. It may reveal more, for example by revealing a selection stage of which the subject was unconscious.) If these requirements are met, then it will be possible to test whether a belief was acquired rationally. There is rational acquisition when a belief is acquired by virtue of a process of reasoning, and in addition the reasons are adequate and are used appropriately in arriving at the belief. The reasoning can be indirect. One might, for example, reason that a given source of information was reliable and that its reliability could easily be checked, so that any drift into unreliability would probably be spotted by someone. One could on that basis accept the deliverances of the source, rather than reasoning about evidence that related directly to specific potential beliefs. We should also note that as with actions, we do not have access to external standards of rationality, but only to the standards of our own community. "Rationally held belief" must therefore in practice be read as "belief that is in our view held rationally". That is the intended reading here, except at the end of this section, where I discuss the notion of standards of rationality that are independent of culture.

I shall limit my attention to beliefs that are of a type that we expect to be acquired following the consideration of evidence. Beliefs of an academic or technical nature are paradigms, but more everyday beliefs, such as a belief that one's friend would arrive late, a belief that was acquired on the basis of a report that train services were interrupted, would be included too. A very wide range of simple everyday beliefs can be tested by considering evidence if we so choose, and to that extent they can be included, although they are not paradigms and it matters little whether they are included or excluded. There are, however, some beliefs that it is not sensible to think of as being acquired or tested by reference to evidence. These are the absolutely obvious facts of life, partly because they are obvious and partly because thinking that they should be acquired or tested by reference to evidence would corrode our way of life, as Wittgenstein argued (*On Certainty*, sections 231 to 236, 253, 310 to 316 and 341 to 344).

I shall refer to beliefs of the type that I have in mind, a type for which academic disciplines provide paradigms, as evidence-basable beliefs. The label is meant to reflect the reasonableness of adducing evidence for the correctness of a belief, rather than the possibility of adducing evidence for its correctness. The class of evidence-basable beliefs includes, but is not limited to, the class of beliefs that was identified in section 1.1, the beliefs that we expect to be adopted following systematic reflection. We only expect systematic reflection when it is reasonable, as well as possible, to consider evidence, and when the beliefs are also sufficiently sophisticated and significant.

The element of justification in the traditional analysis of knowledge has an important role to play here. A subject should be able to justify any evidence-basable belief that she claims to hold rationally. To put the point in the terms that are used by Duncan Pritchard, we expect reflective luck to be minimized in relation to rationally held evidence-basable beliefs, even if it cannot be eliminated (Pritchard, *Epistemic Luck*, section 6.5). That is, we expect someone who rationally holds an evidence-basable belief to believe propositions that are in fact evidence for the belief, and not merely to believe propositions that she takes to be evidence for it. We test for rationality by challenging the belief and considering the subject's responses. The subject should be able to defend the belief in accordance with accepted standards of debate, such as the standards that are catalogued by Alvin Goldman (*Knowledge in a Social World*, sections 5.1 and 5.2). Examples are requirements that the premises that are offered be credible, that they support the conclusion, and that the believer should respond to criticisms in ways that are appropriate both to the seriousness of the criticisms and to the audience's receptivity to available responses. I leave aside for the moment the question of whether justification as such should feature in a definition of knowledge. My concern here is not with the criteria for knowledge, but with the criteria for an evidence-basable belief to be held rationally. I have borrowed Pritchard's analysis of the role of reflective luck for this purpose. He uses it in the analysis of the concept of knowledge and in the analysis of responses to scepticism.

What is needed is not the ability to parrot a file of evidence in support of a belief. Such an ability is not necessary. Someone may, for example,

believe that Descartes was born in 1596, without having any specific evidence in her head. She could still hold the belief rationally. She might remember that she had read the date in a few places, none of which she could now call to mind, and it might occur to her that the writers would have been caught out if the primary sources had not supported their assertions. More significantly, the ability to parrot a file of evidence is not sufficient. Such a file could have been injected into the believer's brain along with the belief that the evidence supported, even if the believer had not considered either the reasons for adopting the belief or the reliability of its source, so that we could not regard the belief as having been acquired rationally.

If something other than the ability to parrot a file of evidence is needed, what is that other thing? We must distinguish between what would be evidence that a belief was held rationally, and what is necessary for a belief to be held rationally. I shall tackle evidence first, and then what is necessary.

There is a straightforward alternative to requiring possession of a file of evidence, which makes the test both more demanding and more reasonable. This is to require an ability to respond to challenges that are drawn from an indefinite range, a range that may not be infinite but that is not delimited in advance. Then a file of responses that was prepared in advance would not be guaranteed to be sufficient, even though it would be a great help. Someone who claims to hold a belief rationally should be able to do more than quote, or gesture in the direction of, a fixed body of evidence. She should be able to respond to challenges like, "What if the evidence were falsified?", or, "Isn't the evidence open to alternative interpretations?", or, "Your claim does not seem to cohere with other claims that are generally accepted to be true". What is required here is the living word that is able to defend itself (Plato, *Phaedrus*, 276a). Responses to challenges would, of course, have to be acceptable. I discuss standards of acceptability later in this section.

It would be too harsh to require someone who claimed to hold a belief rationally to have responses to all likely challenges in a file. This is the respect in which the new test is more reasonable than a demand for a comprehensive file of evidence. The believer that Descartes was born in

1596 who could not produce evidence that would deal with all likely challenges would fail a test that a file should be produced, so that test would be unreasonable. But it would not be unreasonable to require her to come up with responses to the challenges that were actually posed, given time to formulate her responses. Thus the new test of requiring responses to the challenges that were actually posed, those challenges being drawn from an indefinite range, could reasonably be applied to her.

If several challenges to an evidence-basable belief are posed, and the challenges are seen off satisfactorily, that at least shows that the believer is rational not to discard the belief. But the fact that a belief survives challenges is not enough to make it held rationally. It is also important that the belief was acquired rationally in the first place, unless the range of challenges posed is so extensive that the process of responding to them itself involves the thoughtful consideration of a wide range of evidence. Then the process would amount to the rational re-acquisition of a belief that might at first have been a mere guess or a prejudice.

Although the fact that a belief has survived a number of challenges does not guarantee that it is held rationally, we do not need to add a further test. Responses to challenges that are drawn from an indefinite range may themselves provide evidence that a belief was acquired rationally. That evidence will often be good enough. The ability to respond could easily be explained by rational acquisition of the belief in question, for several reasons. First, someone who had reviewed arguments for and against a given proposition would have considered at least some possible challenges to the belief that she finally adopted. Second, if a belief had been acquired rationally, it would have fitted in with the subject's existing beliefs, often in the sense of making her overall picture of the world more coherent and satisfying, a sense that is intermediate between mere consistency and logical implication. That integration with existing beliefs would make it easier for the subject to bring some of those beliefs to bear when responding to challenges. Third, if a subject had considered evidence for a belief and had adopted the belief on the strength of that evidence, the evidence would have satisfied the subject's existing criteria for being good evidence, otherwise she would not have adopted the belief. Furthermore, she would understand why the evidence had satisfied those criteria, giving her further

ammunition when faced with challenges to the belief. So a good performance in a viva voce examination of someone who holds an evidence-basable belief should be taken to indicate that the belief was acquired rationally. Rational acquisition would be the best explanation of a good performance, given that such an examination would probably be failed by someone whose belief had been acquired in other likely ways. Those ways are guesswork, the operation of prejudice, and reading one or two expressions of the proposition that is believed without any consideration of the grounds for believing it.

Responses to a finite sequence of challenges would not show conclusively that a belief had been acquired rationally, because the responses might be provided from a file of responses that had been imbibed without reflection, along with the belief. One would need to enquire into the educational history of the subject in order to rule out that possibility. Even responses to an infinite sequence of challenges would not be conclusive because one might, without reflection, imbibe both a belief and a method of learning from challenges to that belief so as to respond to new challenges, just as a chess-playing computer can develop its talents as it goes along. I shall argue below that we could come to see computers as acquiring beliefs rationally. But that point does not conflict with the point that is made here, because we could see both computers and human beings as having acquired beliefs rationally, even if equivalent performance in viva voce examinations could have been achieved by other means. It would also be important to allow, both in relation to human beings and in relation to computers, that if a method of learning from challenges that was imbibed along with a specific belief was of sufficiently general application, either when it was imbibed or following its development through use, then the fact that it was imbibed without reflection would not matter. If responses to challenges were generated by such a general method, their generation might well amount to more than the resourceful shuffling of a file of evidence. The process of responding to challenges could then amount to the rational re-acquisition of the belief that was being defended. Indeed, the possession of such a method and its use in the defence of beliefs in general would constitute part of the subject's general rationality, a rationality that would increase as the method was developed. The

generality of the method would mean that it was not tied to the belief along with which it had been imbibed. The fact that it had been imbibed along with that belief, as opposed to having been imbibed on some other occasion, would then be of no significance.

On the point that the rational acquirer of a belief would understand why the evidence for the belief satisfied her criteria for being good evidence, it is interesting to note an argument that if we were to take up an internalist position, we could not require someone with justification to be aware of that justification, although we could allow her to be aware of it. The argument is set out by Michael Bergmann ("Reidian Externalism", section 1.1). It is that if awareness of justification were required, a vicious regress would ensue because the requirement for awareness could be re-applied so that the subject would need to believe in the relevance of the justification, and would need to have justification for that belief in its relevance, of which she would have to be aware, and so on. (Bergmann's argument relates to his premise III, on pages 54 and 55. He is concerned with justification for the purposes of claims to knowledge, rather than with the rational acquisition of beliefs.) Bergmann is making a point against internalism. What is interesting here is to see precisely how something external would come in to block such a regress when an evidence-basable belief was at stake. The practice of the community would be the external thing that would underwrite the relevance of justifications for beliefs. It would simply be accepted that some things justified given beliefs. The practice would be a piece of Wittgensteinian bedrock (*Philosophical Investigations*, part 1, section 217, page 72). It is not just a fact, but an important fact, that we accept the role of these shared standards. As Jane Heal has pointed out, we must presuppose shared standards like these if we are to have discussions with a view to reaching conclusions ("Understanding Other Minds from the Inside", pages 43-44).

Beyond the question of evidence for the rational acquisition of beliefs, and arguing for the intrinsic necessity rather than the evidential sufficiency of the ability to respond to challenges, we can claim that the ability to respond to at least some challenges, and to do so to a standard that satisfies our norms, is required by our standards of rationality for evidence-basable beliefs. The starting point of the argument is the fact that if an evidence-

basable belief is to count as held rationally, then the believer should be ready to defend her belief. The mere claim that a believer should be ready to defend her belief does not take for granted the conclusion for which I shall now argue, that she should be able to respond to at least some challenges to an acceptable standard. A defence might not be an acceptable defence. But if someone said, "I believe this, even though I am in no position to defend my belief", that would show that the belief was not held rationally.

A belief may be held rationally even if the first response to a demand for justification is something as feeble as, "Professor Smith told me so", but someone who met our ordinary standards of rationality would be able to add something, for example, that Professor Smith was a renowned expert on the subject. Someone who acquired evidence-basable beliefs rationally would not accept everything that Professor Smith told her, on all topics, unless she was prepared to claim that Professor Smith was a renowned expert on everything, or, more realistically, that Professor Smith was not the sort of person who would pronounce on matters that lay outside his area of expertise. That is, someone who acquired evidence-basable beliefs rationally would sometimes gather evidence herself, and would sometimes rely on authority, but she would not uncritically plug herself into a supposedly authoritative source and allow it to pump beliefs on all topics into her brain. When she had found an expert in a field that was of interest to her, she would not subject each separate pronouncement by the expert to critical examination, but she would be aware that the expert's credentials had to be checked at the outset, and that the extent of those credentials set a limit to the area within which pronouncements did not need to be examined individually. That much is included in our standards of rationality, along with the need to examine evidence and its implications directly when we do not rely on experts. We are not intellectually impressed by someone who does not exercise critical faculties. We may admire people like that for other reasons. No general ethical judgement is intended. Many people do great things because they put their critical faculties to one side and act with confidence on beliefs, the justifications for which they never stop to question. But while they may be praised for their achievements, they cannot be praised for their rationality,

except perhaps in relation to their initial decisions to set their critical faculties to one side in order to achieve as much as possible.

The ability to respond to at least some challenges, and to do so to an acceptable standard, follows directly from meeting our standards of rationality. If a belief has been acquired following critical reflection on the available support for it, the believer will be in a position to respond to challenges in a way that we will recognize as acceptable. (Our generally accepted standards also have a role at the earlier stage of critical reflection. Complicated but meandering and disengaged thought would not put the believer in a position to respond to challenges, but we would not regard such thought as critical reflection.) This does not mean that all challenges will be decisively repelled. To continue the above example, a claim that Professor Smith was a renowned expert might be undermined by a challenge to the effect that his popular reputation was based on work that he had done 30 years ago, and that most experts had now moved on to subscribe to newer theories which implied that Professor Smith's views were mistaken. This challenge might lead the believer to change her mind. The ability to respond to challenges also does not mean that the believer would have either the knowledge or the intellectual horsepower to pursue the discussion of a challenge wherever it might lead. The discussion might get too technical for her. But she would still be able to pursue the discussion for a while, and recognize when and in what respects it was getting beyond her. She could always at least start by saying that her reason for holding the belief was X, and could add that X was a good support for the belief. Her beliefs in the truth and the relevance of X would themselves be rationally held beliefs. The believer would be able to defend them against at least some challenges.

This does not mean that we are threatened either with infinite regress, or with any circularity that should disturb us. We are not here engaged in a foundationalist project to find, or to pursue into the distance, ultimate grounds for all of our beliefs. Instead, we are identifying what is required for the holding of a belief to qualify as rational. It may be enough to go only a few steps beyond the initial belief. Correspondingly, all that has been established here is that if a belief has been acquired rationally, then the subject will be able to offer acceptable responses to some challenges. There

is no implication that the subject will be able to dismiss all challenges decisively. That would be too stringent. But if some significant challenges could not be addressed, it might not be rational to continue to hold the belief.

Rational computers

The uncritical imbibing of a belief, along with a file of evidence for it and a file of responses to possible challenges to it, would not be enough to make the acquisition of the belief rational, and therefore to make the belief rationally held. The obvious analogy is that of loading beliefs, evidence for them and responses to potential challenges into a computer. But any suggestion that computers could do no better than record and repeat what we gave them would do computers an injustice. We must consider what more they could do, whether it might be right to see them as acquiring beliefs rationally, and whether any such possibility should lead us to change our notion of the rational acquisition of beliefs.

It is perfectly possible to supply a computer with a base of information and a set of programs that will use that information creatively, so as to expand the base by taking in new data and drawing inferences. The computer would then be able to respond to questions to which it did not have answers ready in advance. Such a computer could also review existing beliefs, whether acquired as a consequence of the consideration of data or placed in it by its programmers, and could accept, reject or modify them in the light of new evidence or conceptual reflection, just as we can. Expert systems are already with us, and we can expect great progress in this field over the next few decades. A robot has done useful work in genomics (King et al., "The Automation of Science"), and a computer has deduced physical laws of motion from data on actual physical systems, albeit laws that were already known to human beings (Schmidt and Lipson, "Distilling Free-Form Natural Laws from Experimental Data").

Such computers can so far only display impressive capabilities in limited fields. They perform best against human beings in fields that are full of precise definitions and elaborate chains of reasoning, and in fields where repetitive tasks of data collection and analysis must be performed

accurately. But that does not matter. We should ask whether beliefs in some limited field could be regarded as rationally acquired by computers, not whether a computer could pass a test of the rational acquisition of beliefs across all subjects. It also does not matter that the words "belief", "rationality" and "acquisition" might have to be used analogically in relation to computers, although some would maintain that they could be used literally. We do not need to decide that issue here, because the question is that of whether our notion of the rational acquisition of beliefs is open to challenge on the ground that a computer could do the things that we do when we see ourselves as acquiring beliefs rationally, even though it is natural to view computers as not engaging in reflection of the type in which we like to see ourselves as engaging. It will be possible for the challenge to get under way so long as computers could exhibit something that either was, or was analogous to, the rational acquisition of beliefs.

The challenge can indeed get under way. It is entirely conceivable that a computer could respond to an indefinite range of challenges to its beliefs, and that it could do so just as well as a human being. Human consciousness and thought are entirely dependent on brain cells and on the electro-chemical processes that go on inside and between those cells. One physical arrangement could be replaced by another. So a computer could do the same job, although its internal architecture would be different from the brain's architecture, and although it might need to be an analogue computer, rather than a purely digital one (Bains, *Physical computation and embodied artificial intelligence*). The task of the builders and programmers of computers would be made easier by the fact that a computer would not have to give the same responses as a human being would be likely to give. We should not impose the impersonation requirement of the Turing Test (Turing, "Computing Machinery and Intelligence", section 1). The computer would only need to give responses that we, applying our own standards of rationality, saw as acceptable responses, whether or not we would have given the same responses.

If computers could defend their beliefs, we would have to choose between two options. The first would be to amend our picture of the rational acquisition of beliefs so that an ability to defend beliefs against an indefinite range of challenges would no longer be taken to be, on its

own, good evidence of rational acquisition. The second would be to accept that computers could do something that was at least analogous to the rational acquisition of beliefs. We should choose the second option, but we should not fear that we would then have to share our status as human beings with computers. There is a general reason why we would not have to extend to computers our attitude toward human contributors to our debates, whose views we ideally consider on their own merits and do not ignore merely on account of our opinion of the contributors' habits of thought. There is also a reason why, over a wide field, we should not expect computers to be sophisticated enough in their thinking for us to regard their efforts even as decent attempts to acquire beliefs rationally.

The general reason why we would not have to treat the views of computers in the same way that we ideally treat the views of human beings is that the rational acquisition of beliefs is distinct from their being adopted following deliberation with mastery. We can attribute rational acquisition when the process has sufficient sophistication, and when the background speech and conduct of the subject are such that we can interpret what goes on as representing an appropriate path from adequate evidence, via acceptable arguments, to conclusions. We do not need to overlook the causal closure of the physical, nor need we see the process as anything other than the following of a program. But if we do not see a computer as deliberating with mastery, we have no reason to respect its views unless we can identify some other reason to do so. If we think that it is good at reaching correct and useful conclusions, we will generally make use of those conclusions. But its conclusions are evidence for us to use, just like any other evidence. We can take them, leave them or amend them as we wish. If we see something as a mechanism, there is no call to be polite to it, or to ask its permission before changing its programs or the contents of its memory by whatever method is most convenient. There would be no need to limit ourselves to methods that amounted to, or that resembled, rational persuasion. It would not be rational for us to contradict the output of a reliable computer on a topic that mattered to us, unless we thought that it had not taken all significant considerations into account. But if we did contradict it, that would be a failure of rationality in ourselves, not a gesture of disrespect to the computer.

The reason why we should, over a wide field, not expect to have to

regard computers' efforts even as decent attempts to acquire beliefs rationally, is that they would lack the understanding that is needed in order to say much that is worthwhile in fields such as the humanities and practical ethics and politics. On the one hand, computers can behave like us within a limited but broadening field, and can out-perform us as rational acquirers of beliefs in some parts of that field. On the other hand, computers lack the experience of life, and the consequent interplay of reasons, emotions and values, that we recognize in one another, a recognition that persuades us to take note of one another's views in the humanities and in practical ethics and politics. The lack of experience would obviously matter because it would leave open the way to logically impeccable but unwise conclusions in practical life. But it would also matter in relation to purely academic beliefs, on which no-one would act beyond writing the next round of academic papers. Some propositions can only be understood sufficiently well to acquire them thoughtfully, or to defend them, if one has appropriate experience of life.

The relevance of such experience depends both on the topic and on the style of the proposition. For propositions within mathematics and the natural sciences, the relevance is nil or very small. The relevance is also nil or very small for propositions within the humanities that are of a very straightforward factual style, such as propositions that give the dates of historical events. The relevance gradually grows as we move up to more subtle propositions. Suppose that John asserts that Napoleon was in complete command of his armies. He is challenged by Jane, who has read the epilogue to *War and Peace* and who shares Tolstoy's view that Napoleon's command was largely an illusion, depending as it did on whether people whose conduct was outside his control acted in ways that happened to correspond to his orders. John could only respond in a way that we would recognize as satisfactory if his experience of life had been rich enough to generate an appropriate interplay of reasons, emotions and values within him. He would need to know what it was to be in command, what it was to be obedient or wilful, and what it was to pursue one's existing course regardless of orders coming down from on high. John would need to have at least some empathetic understanding of Napoleon, of the various ranks of officer under his command and of ordinary

soldiers, in order to use the historical evidence judiciously. Computers would be most unlikely to be capable of this. To take an example that would be even more likely to expose the limitations of computers, the claim to be defended against challenges might be that the chief lesson that we moderns should take from Pindar's odes was that of the importance and the fragility of personal triumph.

It does not follow that we could only see a computer as having acquired beliefs rationally if the beliefs concerned the subject matter of mathematics and of the natural sciences, or if they were straightforwardly factual beliefs of types that could be recorded in catalogues. Rather, there are degrees. A suitably informed and sophisticated computer could conduct some respectable arguments about questions in the humanities. It is just that it would fall short of what we could manage, both in the range of arguments that it could conduct and in their quality.

The fact that computers would need to have certain experiences in order to function well in some fields points to a reason why it might be appropriate to contradict the conclusions of a reliable computer, if the topic was of a certain type. We might say that the computer did not have the understanding to make wise ethical or political decisions, or to reach sensible conclusions in the humanities. This would be a different reason from the one noted above, that we might rationally contradict the conclusions of a reliable computer if we thought that it had not taken all significant considerations into account. A lack of understanding is not the same thing as the overlooking of a specific consideration, although the former can very easily lead to the latter. Correspondingly, there is a range of different reasons why it might be appropriate to refuse to regard a computer, or a human being, as having rationally acquired an evidence-basable belief. What matters for rational acquisition is that the belief should be appropriately related to the evidence, with the relation being given by an argument that is generated by the use of an appropriate method. A failure to take all significant considerations into account would be a failure at the stage of the weighting of evidence. A failure of understanding could lead to the same type of failure, but it could also lead to a failure to use an appropriate method of argument, not because the subject chose the wrong method from the range that was available to him, but because the right

method was not available to him. This would not be a culpable failure, nor would its occurrence cast doubt on any claim that might be made to see a subject as deliberating with mastery. But it would still prevent us from seeing the belief in question as having been acquired rationally.

The strength of the argument from computers' lack of experience of life, and from their lack of our kind of interplay of reasons, emotions and values, is contingent on the state of technology. Such deficiencies may well be made good as computer science develops, especially if computers are made mobile so that they can explore the world, and if they are put through sequences of experiences that are analogous to growing up, rather than springing fully formed from the factory. Alternatively, future computers might go through different developmental processes, leading to new types of difference from us. They might come to be on a par with us, but we might not be able to appreciate that they were on a par with us, or be able to see them as rational by our lights. They might come to have their own questions in subjects that were analogous to the humanities. They, but not we, might then be able to arrive at answers to those questions in ways that were, by their lights, rational.

I shall now discuss what would constitute acceptable responses to challenges, before turning to the long-standing problem of scepticism.

Acceptable responses to challenges

If an ability to respond acceptably to challenges that are drawn from an indefinite range is of interest as good evidence of the rational acquisition of at least some beliefs, it will be useful to understand the criteria for responses to challenges to count as acceptable. It is not enough that responses to a sequence of challenges be understood by the person who hears them, or that they be recognized by that person as responses. They must be sufficient to ward off challenges, although it may still be reasonable for the challenger to persist by mounting fresh challenges. I shall take responses to be acceptable if they meet this modest standard of warding off challenges. If we only seek evidence of the rational acquisition of beliefs, a demand that every challenge be decisively defeated would be too stringent.

In practice, we rely on the consensus of relevant members of our society to give us standards of acceptability. But we can ask whether that

is enough. Would a majority vote as to the acceptability of a response, even a decisive majority of the experts in the relevant discipline, or a decisive majority of all of us when the belief was not of a specialist nature, really prove anything? What if we all used a mistaken standard of acceptability of responses to challenges? An alternative worry is not that there is some correct standard that we fail to grasp, but that there is no such standard at all, and that the existence of a consensus would demonstrate nothing because there was nothing to demonstrate. This latter worry would prompt the fear that we should see a serious threat to our standards of rationality, rather than an obvious truth about what we can access, in the claim that "there are no context-free or super-cultural norms of rationality" (Barnes and Bloor, "Relativism, Rationalism and the Sociology of Knowledge", page 27). That would be an unnerving prospect for those who, like me, believe both that there are, in a straightforward sense, true and false propositions, and that we can act sensibly to discover which propositions are true.

Fortunately, the threat can be addressed. We can see this by recognizing that our standards of acceptability of responses are derived from our standards of rationality in the acquisition of beliefs, and that compliance with those latter standards is conducive to the acquisition of true beliefs. The first stage in the argument is to note that responses to challenges will generally be thought to be acceptable when the responses are such as to disclose that rational acquisition took place. The responses will be intended to be defences of the belief, rather than defences of the claim that the belief was acquired rationally, but the best way to defend a belief is to show how the proposition that is believed follows from some evidence. In so doing, the respondent will indicate that the belief was acquired rationally, because she will show a grasp of how one could arrive at the proposition by applying an accepted method of argument to some evidence. She might not have acquired the belief by that route, but her knowledge of the route would reasonably incline us to the view that she had taken it.

The second stage in the argument is to note that we do not seek to acquire beliefs rationally purely for the pleasure of exercising our grey cells. We seek true beliefs. We have discovered that the activities of weighing

evidence and of choosing our methods in accordance with accepted standards, activities that we regard as rational when so conducted, are conducive to the acquisition of true beliefs. There is a link between the rational acquisition of beliefs and their truth, but it is not a perfect link, and the truth of a belief is not constituted by its mode of acquisition. If we rely on consensus when stating what constitutes the rational acquisition of a belief, or when stating how to recognize rational acquisition, and if we thereby rely on consensus to determine standards of acceptability of responses to challenges, that does not imply the reduction of truth to a majority vote, or even to a unanimous vote. There may not be any norms of rationality that are independent of culture. But that would not prevent the existence of truth that was independent of culture, nor would it block our access to such truth. Putting the stages of the argument together, compliance with our consensual standards of acceptability of responses to challenges indicates compliance with our consensual standards in the acquisition of beliefs, and that latter compliance allows us to increase our chances of coming to believe propositions that are straightforwardly true. References in the course of this argument to true beliefs do not allow the relativist to object that the argument is circular, because the aim is only to show that the notion of straightforwardly true beliefs can be accommodated in the absence of super-cultural norms of rationality. The aim of the argument is not to refute relativism in general.

We can also take comfort from the fact that it may be possible to say something that is independent of culture about the structure of standards of rationality, even if not about their content. An example is given by Michael Smith, in "The Structure of Orthonomy". That paper is concerned with rational action, and specifically with rational action that springs primarily from the confluence of beliefs and desires, but one could construct something similar specifically for adoptions of belief, where the most natural desire would be to adopt true beliefs and to reject false ones. In addition, we can take comfort from the arguments that are deployed by Paul Boghossian to show that we could justify our own epistemic system over at least some alternatives, if only by our own lights (*Fear of Knowledge*, chapter 7). Boghossian also shows that even if relativism did get a toehold, the impact of any plausible relativism would be severely

limited. In particular, it is not plausible to claim that facts about the world depend on our thoughts, in any way that should disturb us (ibid., chapter 3). More generally, we need not go so far as to abandon claims that some propositions are definitely mistaken and that others are almost certainly correct. We do not have to concede equal validity to different ways of looking at the world. Perhaps we do exalt our own knowledge, but we are not necessarily wrong to exalt it.

The only way in which we can actually test the acceptability of responses to challenges is to take the opinions of people, possibly a select group of experts or possibly the whole population, and to combine those opinions in various ways, such as a simple vote or a weighted vote. In highly technical fields, we might do something sophisticated. If, for example, a new scientific theory were subject to challenge and its proponents had offered responses, we might call an academic conference at which experts who had not already taken sides would work through the arguments and would debate whether either the challenges or the responses missed the point. But we would still be taking people's opinions, not referring to an independent standard. There is no other method of testing, and there may not be any other criterion of acceptability, even an inaccessible one (assuming that the concept of an unusable criterion might be coherent). But given that this does not put truth that is independent of culture out of bounds, I am content to rest with a consensual standard of acceptability of responses. This may not be ideal, but it is as good as anything that we can do, and that may be so for deeper reasons than the practical one that we can only ask people around us for advice. To go back from standards of acceptability of responses to the standards of rationality in the acquisition of beliefs that give rise to them, Timothy Williamson offers an independent argument for the conclusion that we may not know what we need to do in order to act rationally in our search for knowledge (*Knowledge and its Limits*, section 8.7). In his words, "Rationality may be a matter of doing the best one can with what one has, but one cannot always know what one has, or whether one has done the best one can with it" (ibid., page 179). Williamson does not, however, counsel despair, any more than I do. As he points out, we can usually manage pretty well, even within our limits.

My arguments here do not imply a position either for or against the

existence of verification-transcendent truth. The identification of beliefs that have been, and of those that have not been, acquired rationally does not determine which beliefs are true, although the process of challenge and response helps us to sort truth from falsehood. The concept of the rational acquisition of a belief might very well be limited to what we could determine, so that rationality of acquisition could not be verification-transcendent. But the fact that truth was a separate matter from rationality of acquisition would allow, although it would not imply, that truth could go beyond any limits that applied to rationality and could be verification-transcendent.

There is a positive aspect to the fact that the opinions of those around us have a vital role to play in the formulation of norms of rationality in the acquisition of beliefs. The foregoing remarks suggest that it is palatable to abandon an aspiration to universal standards and to be pulled down to culture-bound standards. But we also need to be pulled up from individual standards, standards that could all too easily be quirky and unstable. We need help from people around us. I would not go so far as to claim with confidence that a lifelong solitary would be debarred from having any standards of rationality. The pains and joys that he experienced, as a result of his attempts to navigate through the world and to manipulate the world, might provide enough feedback to make it clear to him whether he was thinking and acting appropriately. (One would need to allow him to have self-consciousness in order to make this possible, in defiance of Nietzsche's argument that he would not even have reflexive consciousness because he would have no need of it (*Die Fröhliche Wissenschaft*, section 354). He might also not formulate the concept of rationality, because he would never have his rationality challenged by others.) The necessary feedback might even be available in relation to abstract thoughts that had nothing to do with his daily life. He might, for example, recognize the thrill of success that would come from formulating a concept that unified several areas of mathematics, and he would rightly conclude that he was conducting his mathematical arguments in an appropriate way. But if the arguments against semantic individualism in Sanford Goldberg's *Anti-Individualism* are to be accepted, it is doubtful whether a lifelong solitary could engage in an extended and successful chain of rational abstract thought. Success

would depend on the transmission of information from the solitary at one time to himself at a later time, and there would be no social calibration to keep the mechanism of transmission in good working order. A. C. Grayling would even deny such a lifelong solitary a language, despite the fact that any such language would only be contingently private. Grayling's reason is that speaking a language requires following rules, and that speakers must have the means to distinguish between following the rules and only thinking that they are doing so (*Truth, Meaning and Realism*, pages 83-85). A parallel argument for standards of rationality would be that a lifelong solitary could not possess articulated standards that he could apply to his own processes of thought independently of his feeling that a particular train of thought had come out well, or his feeling that his planned manipulations of the world had enhanced his comfort in the ways that he had desired. Even without going that far, and even allowing that a lifelong solitary might possess articulated standards of thought that we would recognize to be appropriate to such an extent that we would call thought that accorded with those standards rational, we can see that he could all too easily wander off track, cease to meet the standards with which he had started, and not notice that he had so ceased. Likewise, we non-solitaries require regular input from others in order to be confident that we are staying on track.

5.4 Scepticism

In this section, I shall move back from evidence-basable beliefs in general to the class of beliefs that was identified in section 1.1, the beliefs that we expect to be adopted following systematic reflection. I shall ask what the sceptic can do to unnerve us as to whether we have enough justification to make such beliefs candidates for the status of knowledge. I shall not establish what other conditions there might be for beliefs to qualify as knowledge, or whether it might be feasible for our beliefs to meet those other conditions. There are those who think that we should concentrate on justified belief anyway, and that we should not worry about knowledge as something beyond that (Kaplan, "It's Not What You Know that

Counts"). As I concentrate on the sceptic about justification, my remarks are just as relevant to that view as they are to the view that knowledge is what matters but that justification has a lot to do with knowledge. One other point is worth making in order to locate my remarks in relation to the work of others. I shall make considerable use of our standards of rational acquisition of beliefs. As we have seen, the ability to respond to challenges can be used as evidence of rational acquisition. That ability can also be used in the analysis of knowledge itself. This is demonstrated by Robert Brandom's use of his default-and-challenge structure of entitlement (*Making It Explicit*, pages 176-178 and chapter 4; see also the use that Michael Williams makes of that approach in "Responsibility and Reliability").

I shall start by setting on one side any demand for certainty. I shall then touch on the relevance of justification, and outline the structural attack on the worth of our justifications, the attack that is summarized in Agrippa's trilemma. The trilemma is only meant to be a decisive argument against the possession of knowledge if knowledge requires legitimate certainty, but even if we do not expect certainty, the trilemma may still have a power to unnerve us. The next stage is to argue that our standards of rational acquisition of beliefs can themselves be a source of justification for beliefs that have been acquired in accordance with them, and that the sceptic about justification who wishes to unnerve us must therefore attack the justificatory value of compliance with those standards.

The next stage is to argue that the sceptic must mount his attack in a purely structural form. I first consider attacks on specific types of justification for specific beliefs, and argue that if such local scepticism is to be truly local, and if we do not expect certainty, it cannot be scepticism of a philosophically interesting sort. I argue that this conclusion holds even if types of justification are defined very broadly. An example is the type, sensory perception. The sceptic is therefore driven back to the purely structural form of his argument. Furthermore, the sceptic cannot use that argument to attack the worth of the structured sets of justifications that we offer for specific beliefs. The attack must be on the justificatory value of our general methods of arriving at beliefs. Here the sceptic fails because he takes too narrow a view of justificatory structures. In some disciplines, the

predominant method is to build up theories from axioms. In others, it is to argue in large circles. The use of either method can confer perfectly good justification. Circular arguments can do so even though there may be equally strong arguments for contrary conclusions, and constructions that are founded on axioms can do so even though the axioms are unsupported.

It is important throughout this section to remember that rational acquisition of the type of belief with which we are concerned is not merely a matter of having an idea. A belief is only acquired rationally if the subject has duly considered the evidence for and against the proposition that is to be believed. That will always require the construction of arguments, and it will often require active testing. So rational acquisition may in practice follow some time after acquisition.

The sceptic and certainty

The sceptic is the perpetual bugbear of epistemology. Whenever we define knowledge in a way that is commensurate with the status that it has for us, we impose conditions on a belief's being knowledge that make it all too easy for the sceptic to attack the claim that any given belief counts as knowledge. Sceptical attacks may fail, but we have a hard time fighting them off and even then, the sceptic is soon back with another line of attack. We can defeat the sceptic more decisively by defining knowledge in less demanding ways, but then we run the risk of using a concept that is too weak to correspond to our pre-philosophical notion of knowledge.

One line of sceptical attack is to point out that we cannot be certain of the truth of any of our beliefs. There is always something that could have gone wrong. Even confidence in one's mathematical beliefs can be undermined by the thought that one's mind might have wandered when devising or reading proofs, so that mistakes might have escaped one's attention. Once sensory data are involved, the possibilities for error multiply. It is stretching credulity to think that all those who have ever studied Pythagoras's theorem, or all those who have ever observed that elephants are larger than lions, have been mistaken. But the sceptic can suggest that his interlocutor might be a brain in a vat, systematically deceived about the whole world.

The standard, and in my view appropriate, response to this line of

sceptical argument is to deny that knowledge only exists when the knower can exclude the possibility of error, while accepting that there must be no actual error. Those who insist that knowledge must be closed under entailment, so that if I know that I am now at my desk I must know that I am not a brain in a vat, are essentially re-packaging the view that the knower must be able to exclude the possibility of error, although as Duncan Pritchard points out, a principle that knowledge is closed under known entailment is logically weaker than a principle that a knower must know that he is not in error (*Epistemic Luck*, page 27). The denial of the need for the knower to be able to exclude the possibility of error may be formulated directly or indirectly. An example of the latter is formulation in terms of a requirement only to consider epistemically relevant worlds, worlds that are near enough to the world as we think it actually is (DeRose, "Solving the Skeptical Problem", sections 11 and 12). We should also note the difference between certainty and the impossibility of error. Certainty requires belief in the impossibility of error on the given occasion, and legitimate certainty requires the actual impossibility of error on the given occasion, but the impossibility of error does not require anyone to be certain of anything. Certainty would only arise if someone considered whether error was possible, concluded that it was not, and recognized the implications of that conclusion.

Justification scepticism

Another line of sceptical attack is to put pressure on the justifications that we offer for the beliefs that we claim to be knowledge. No matter how the traditional definition of knowledge as justified true belief might best be modified, it is likely to include something along the lines of justification. It is only likely, and not certain, because there are arguments for a purely externalist and reliabilist concept of knowledge, one that would attribute knowledge to anything that absorbed and mechanically processed environmental influences in ways that tended to improve its performance in specified tasks, such as coping with threats or supplying accurate information on demand. But such an extreme concept of knowledge is not guaranteed to win the day. It clashes with our intuition that knowledge, and certainly knowledge of a conceptually complex nature, should have been acquired thoughtfully. If we consider other philosophical positions that limit

the role of justification, we find that even Timothy Williamson accepts that justification can be significant, despite his argument that knowledge is not to be analysed into components along the lines of justified true belief, and despite his proposal to invert the analysis by equating a subject's evidence with his knowledge (*Knowledge and its Limits*, section 1.3, page 41 and chapter 9). Even William Alston, when he argues that we should discard the concept of justification, replaces it with a range of epistemic desiderata, several of which have a distinct aroma of justification (*Beyond "Justification"*). I shall therefore take it that justification, or something very similar to it, is important. If that is so, then the sceptic has a line of attack. I shall call the sceptic who pursues this line, the justification sceptic. (Michael Williams uses the term "Agrippan sceptic" in chapter 5 of *Problems of Knowledge*, but I shall range beyond Agrippa's trilemma.)

A belief may be accompanied by a structured set of justifications. There will be some justifications that directly support the belief, and there may be others that support belief in those immediate justifications, or in their relevance. The structure may continue through several levels of support. The sceptic's structural argument proceeds by identifying and criticizing the possible structures. Do the immediate justifications for a belief themselves have justifications, or not? If they do, are there justifications below them, in an infinite downward tree, which would hardly seem to be satisfactory? Or are there justifications at some basic level, which do not themselves require justification? That would also seem to be unsatisfactory, both in general because we should not be asked to take anything on trust, and because specific examples of putative basic justifications, such as beliefs that correspond to the direct deliverances of our senses, could be mistaken. Another possibility would be for all of our beliefs to be linked in a vast web of mutual justification, in which some beliefs supported others but nothing supported the whole web. That too would be an unnerving prospect. We would have only local, and not global, justification. We could only hope that the whole web was not a large and internally coherent mistake. Finally, if the immediate justifications for a belief do not themselves have justifications, and if they are not basic justifications that we can reasonably take on trust, how can we really have any justification at all?

The options are neatly summarized in Agrippa's trilemma. We have a choice between infinite chains of justifications, mutually supporting justifications and dead ends in chains of justifications. The trilemma only purports to be a decisive refutation of any legitimate claim to certainty that is based on justifications. But it looks unnerving even if we do not claim certainty, but limit ourselves to claiming that some of our beliefs are correct and that some of them, perhaps not all or only the correct ones, have a reasonable amount of justification. If justifications do not legitimately come to an end, or if they are ultimately only mutually supporting, can they do anything much to engender confidence in our beliefs? In the remainder of this section, I shall argue that the justification sceptic cannot unnerve us in this way, so long as we do not expect certainty.

The rational acquisition of beliefs

The concept of a rationally acquired belief can help us to respond to the justification sceptic. We recognize that some beliefs have been acquired rationally. If we are in doubt about a belief that someone holds, we can test for rational acquisition by challenging the belief and seeing what responses are offered. But normally, we do not feel the need to check. We are aware of how we think and of how we acquire beliefs, although the details of the brain's workings, and some of our biases, are hidden from us. We know that we reach our conclusions on the basis of reasons, and while we know that our reason can be subverted by propaganda and by advertising, we regard such subversion as a failing that is to be avoided so far as possible. Each of us has a sense that other people likewise reach their conclusions on the basis of reasons. One person's recognition that another has acquired his beliefs rationally is a result of her grasping that he has proceeded in a way in which she might also have proceeded, even though she might have reached different conclusions.

Our standards of the rational acquisition of beliefs are not purely individual. As noted in section 5.3, there are social standards. Individuals apply the notion of rationality, but we do not have to accept the way in which any given individual applies it. Most of us take it as part of rationality for each person to reflect carefully on his own views, including his views on rationality, if they appear to differ from those of the vast majority, not in

order to submit uncritically to the majority but in order to consider whether he might be mistaken. This, and rather more significantly the fact that our understanding of rationality is shaped by an education to which many people contribute, either directly as parents, teachers and the like or indirectly through contributions to the culture in which we live, mean that fairly stable standards of rationality are shared by a large number of people.

The fact that the criteria by which we recognize rational acquisition are not primarily individual, but are largely social, gives us confidence that our own standards of rationality are not mere personal quirks. This confidence allows us to apply our standards to ourselves, with some confidence that we are not simply licensing any old way of proceeding. That in turn allows us to claim that we have justification for beliefs that are regarded as acquired rationally.

We can claim to have justification because the notions of rational acquisition and of justification are two sides of the same coin. We regard each belief as acquired rationally if and only if we consider that we had decent justification for it when we acquired it, whether justification that related directly to the specific belief, or the indirect justification that would come from the belief's having been acquired from a reliable source. (When justification evaporates, because new evidence comes to light or because a defect in the original procedure is noticed, we do not re-write history and say that the belief was not acquired rationally, but we may regard it as irrational to continue to hold the belief.) Furthermore, the connection between rational acquisition and justification is a connection of substance, not a mere linguistic habit. The justification that is associated with rational acquisition is worth something because there is a positive, if imperfect, correlation between beliefs' being regarded as acquired rationally and their being true. We can also note a connection between rational acquisition and the value of knowledge. The rational acquisition of true beliefs confers resilience on those beliefs. That is, a believer who has thought about the evidence is less likely than one who has acquired the same belief non-rationally to be led to abandon the belief by misleading counter-evidence. Miranda Fricker identifies this resilience both as a typical feature of knowledge, and as central to the value of knowledge ("The Value of Knowledge and the Test of Time").

Beliefs that are regarded as acquired rationally are not guaranteed to be true, but our experience is that we do quite well. The sceptic can counter that we only have inductive reassurance, and that standards of rational acquisition that have worked well so far might start to lead us astray tomorrow, but we can respond that this would not show that our standards had not worked reasonably well so far. The sceptic can also counter that any reassurance that we may have is based on our fallible interpretations of what has happened so far. We only think that we have mostly done quite well, and we might be wrong in so thinking. But this would be an attack based on a supposed need for us to have certainty, not directly at the level of our first-order beliefs about the world, but directly at the level of our higher-order beliefs about the usefulness of our standards, and therefore indirectly at the level of those of our first-order beliefs about what had happened that supported those higher-order beliefs.

The sceptic who attacks our claims to have justification for our beliefs, but who concedes that legitimate certainty is not essential for knowledge to exist, therefore needs to attack the value of our use of the methods by which we acquire beliefs. But we must first explore other lines of attack, in order to show that the sceptic cannot get far either by attacking specific types of justification, or by attacking the structured sets of justifications for specific beliefs.

Specific types of justification

It might seem that the justification sceptic could win by concentrating on the justificatory power of specific justifications for specific beliefs. The sceptic can certainly argue for local scepticism by challenging specified types of justification and showing, for example, that our habitual reliance on certain scientific instruments is suspect. But such attacks are effectively demands for a demonstration of the impossibility of error, except to the extent that they are contributions to the improvement of our apparatus for collecting and analysing data, or contributions to our understanding of the limitations of our apparatus. That would not be scepticism of a philosophically interesting sort. Thus a sceptic about the results of an experiment in physics might point out that the instruments that took measurements could have been affected by electromagnetic fields that were

generated by other equipment, or that the instruments were only accurate to within five milliamps. The appropriate response would be to turn off the other equipment, or to include appropriate error ranges in the statement of results, not to send for a philosopher. The philosophical sceptic is not interested in that sort of case. He argues that our senses, or our scientific instruments, or our faculties of mathematical reasoning, might be systematically misleading, or occasionally but undetectably faulty, as opposed to being occasionally and detectably faulty, or detectably limited in their accuracy. The local sceptic would have us take such worries seriously, and he can only make us do so, while remaining local in his scepticism, if he first convinces us that knowledge requires legitimate certainty.

We can take the argument against local scepticism further. The sceptic could not in general succeed if he attacked specific types of justification, even if those types were defined in very broad terms, such as the type, sensory perception. The argument here will show that it is no accident that local scepticism is philosophically uninteresting if we have no expectation of certainty, even when the localities in question are made very large by defining types of justification very broadly. Suppose that the sceptic challenged the use of justifications of type 1 (a set of justifications considered together). We could then step outside that type of justification and support the use of members of that set by citing other justifications of types that were not under challenge, justifications of type 2 and of type 3 (a specific member or members of each set). For example, the use of justifications of the sensory perception type (type 1) could be supported by specific results from physics (type 2), which explained the transmission of light, sound and other signals, and by specific results from human biology (type 3), which explained how our sensory organs and our brains worked. The justification sceptic could not simply respond by widening the type under attack, type 1, to make it type 1-or-2-or-3, because the use of justifications of types 2 or 3 (as whole sets) could be supported by specific justifications of types 4 or 5 respectively, and the use of justifications of those types (taken as whole sets) could in turn be supported by specific justifications of other types. To pursue the example, the use of justifications that were drawn from physics might be supported by the coherence and fertility of its mathematical formulation (a justification of

type 4), and the use of justifications that were drawn from human biology might be supported by its success in helping us to devise treatments for diseases (a justification of type 5). The debate would then move on to the legitimacy of using justifications that were based on patterns in our abstract reasoning (type 4), and to the legitimacy of using justifications that were based on practical success (type 5).

The justificatory interdependences between types of justification and individual justifications of other types mean that there are many ways of justifying the use of justifications of a given type. We may expect that the justification sceptic who started by attacking the use of a specific type of justification to support a given belief would repeatedly find himself outmanoeuvred by the defender of the belief, although the sceptic might well be able to mount a fresh attack following each manoeuvre. I deny that the sceptic could in general define a class of types of justification, such that if he could attack the legitimacy of using justifications of each type, the defender of the belief would have no further way of justifying his belief. The sceptic would have to identify a class such that each of the available sufficient structured sets of justifications for the belief included at least one justification of a type that fell within the class, and had to include that justification in order for the structured set to be sufficient. There might be particular cases in which the sceptic could do that, but across most of our knowledge, it looks unlikely that he could do so. The onus is on anyone who wishes to maintain a general scepticism to show that it could be done.

My claim would be undermined if structured sets of justifications were expected to guarantee the truth of the beliefs that they justified. If a guarantee of truth were expected, it would be relatively easy to identify specific essential underpinning justifications, such as justifications for ignoring the possibility that one was a brain in a vat, or essential types of justification, and to attack those essential justifications or the use of justifications of those essential types. But I have already rejected the demand for a guarantee of truth. And my claim will be enough for my purpose, which is to drive the sceptic back to use of the purely structural form of his argument. My argument does that because it shows that sceptical attacks on the value of specified types of justification can in general be expected to be open to counter-attacks. It is enough merely to

drive the users of those attacks back to the purely structural form of the sceptical argument, because if a sceptical attack is to convince, it must be conducted against those of us who think that we have knowledge. There is no need to drive anyone else back to the purely structural form, because only the user of an argument against the value of specified types of justification can engage with us. We could ignore anyone who claimed that there existed some argument against the value of unspecified types of justification that would undermine our supposed justifications. Arguments against the value of unspecified types of justification cannot convince in epistemology, because it is not a subject in which everything is defined precisely in advance within a fixed structure. It is not like logic, in which we can rightly be convinced by arguments of comparable generality, for example, arguments to show that all consistent logical systems of a certain level of sophistication are incomplete. If types of individual justification are not specified, then the sceptical argument cannot convince if it is an argument against the value of individual justifications of all types. The sceptic cannot rely on universal quantification over types of justification. He must abstract from those types and give an argument against the value of structures.

There may be more than one purely structural argument that the justification sceptic could use. He might not have to argue that every structured set of justifications suffered from the same structural defect, such as the defect of not terminating in self-justifying justifications. (All three of the types of structure that are identified by Agrippa's trilemma suffer from this defect. Infinite chains and mutual support suffer from it because chains of justifications do not terminate at all.) He might be able to categorize structures of justifications, and then assign different defects to structures of different categories. But any such argument would still be purely structural, in the sense that the categories would not be categories of individual justifications, such as the category of sensory evidence or the category of results that were drawn from biology. If the categories were of that nature, then the sceptic's argument would be vulnerable to the type 1, type 2, type 3 argument that was deployed above.

Sets of justifications for specific beliefs

Having pushed the justification sceptic back to the purely structural form

of his argument, I shall now argue that he cannot use attacks on structured sets of justifications that are specific to given beliefs in order to establish global scepticism, in the limited form of unnerving those of us who reject the need for immunity to error. He will then have to fall back on attacking the general methods by which we arrive at beliefs that we regard as acquired rationally. He will have to try to show that those methods do not confer enough justification to give the beliefs a chance of being legitimately regarded as knowledge. Then I shall argue that he cannot succeed even in that attack.

The justification sceptic cannot achieve his goal by attacking the value of structured sets of justifications that are specific to given beliefs, because while the standard of acceptability that he imposes on structures of justifications would be appropriate if the role of justification were to guarantee truth, it is too demanding for justification's role in giving us as much confidence in our beliefs as we can reasonably expect. There is the negative point that it is not rational to expend disproportionate effort on establishing the truth or falsity of propositions that have only limited importance. That is not a very appealing point in isolation, because it would be consistent with our capabilities being so limited that overwhelming ignorance was our inevitable lot. But we can add the positive point that we can have a fair amount of confidence in the outcomes of our investigations of the world, so long as we do a reasonable amount of work to establish the truth or falsity of each proposition with which we are concerned. Our position is particularly strong when we justify our use of some types of justification by reference to justifications of different types. We may, for example, have, as our primary justification for a belief, some output from an electron microscope. We may then justify both our reliance on the use of that instrument, and our ways of interpreting its outputs, by reference to independent physical theories.

Justification is cast in the role of a guarantor of truth in Scott Sturgeon's paper, "The Gettier Problem". But that role is only a legitimate one to impose on justification if immunity to error is a legitimate requirement for knowledge. I have already rejected that demand. Sturgeon imposes the role in response to Gettier cases. He does not do so as an ad hoc solution, but on the basis that Gettier's problem reveals something deep

about the nature of knowledge. Even so, it is not at all clear that this is the best way forward, once we weigh up all considerations. The imposition of a condition of immunity to error would, as Sturgeon notes, leave us with very little that would count as knowledge. That might turn out to be correct. Our intuition that we know a great many things might need to be discarded. But that intuition is very strong, and it should not be discarded lightly.

Our methods of acquiring beliefs rationally

Moving on from the argument as it applies to structured sets of justifications that are specific to given beliefs, the justification sceptic can attack the general methods by which we arrive at beliefs. This is where he fails directly, rather than failing by arguing that we lack legitimate certainty when that is not needed anyway. He points out that we have to choose between an infinite structure, a structure that relies on mutual support and a structure with unsupported foundations, but structures of the last two types can confer substantial justification. We can still be concerned when we notice such structures in specific cases. Some people are disturbed when they learn that mathematics is built on axioms that are simply accepted. Bertrand Russell, on being told at the age of 11 that he just had to accept Euclid's axioms, saw his hopes of finding certainty crumble ("Why I Took to Philosophy", page 57). We do not accept circular arguments as deductive proofs of the truth of their conclusions. And we can be unnerved when we notice circles in our arguments, even when we do not seek deductive proofs. But such concerns do not prevent us from accepting that significant justification is conferred by the methods that we use, even though the use of those methods creates sets of justifications that the sceptic would regard as having objectionable structures.

To start with axiom-based knowledge, we accept mathematical reasoning as perfectly adequate to justify belief in its conclusions. Ample justification is given by the huge success of mathematics, as displayed in the ways in which the discipline has unified large areas of our knowledge and in the stunning structures that have been built on a base of simple axioms. The tightly-knit logical structure of mathematics is a source of its fertility, because it allows concepts and techniques from some areas to be used in

what might at first appear to be unrelated areas. In physics too, mathematical axioms and tight logical integration are conspicuous. It is characteristic of physics to seek to explain everything within its scope in terms of the most fundamental particles and forces, and the most general characteristics of space-time. It is, however, true that observations become just as important as axioms as soon as we step from mathematics into physics. We need observations in order to select between possible mathematical models, and in order to fix the values of parameters. And the massive empirical success of physics is a vital source of justification, alongside the fact that our physical theories are given in tightly-structured mathematical forms.

In both mathematics and physics, we expect to show how the immensely complex arises out of the very simple. It is not surprising that we should end up with unsupported foundations, a structure to which the justification sceptic objects. Indeed, if we are to get down to basics, they have to be just that, basics, with nothing as yet found underneath, although we may in due course dig deeper and find new basics. Our justification for our selection of foundations is to be found not by looking further down, but by looking up at the magnificent, and hugely successful, structures that we have built on those foundations. If the justification sceptic will concede that there may be some set of foundations that is correct, we can at least say that there is no sign that we have made a foolish selection, even if the history of science, with the fundamental changes of theory that have taken place, may limit our confidence that our current selection is correct. (If no set of foundations could be correct, any selection could only be pragmatically acceptable.)

At the opposite extreme, arguments in the humanities, and practical arguments about ethical and political questions, rest on hardly any axiomatic foundations, and may very well turn out to be circular if one analyses them carefully enough. (I refer to circular arguments because although the relationships of support that exist in a full set of beliefs have the form of a complex web, the arguments that are actually expressed for specific conclusions have far less complexity of structure, even if their structures are not quite as simple as mere circles.) We still regard such arguments as justifying belief in their conclusions, both because the circles

are large and bring many considerations to bear, and because the conclusions really do deepen our understanding of the human world. The role of circles is one reason why a Verstehen approach looks appealing, although the use of such an approach is neither required, nor to any more than a modest extent explained, by the fact that circles play a role. The picture that I am painting is not intended either to posit or to supplant a distinction between Erklären, explanation, and Verstehen, understanding.

There are many disciplines in between the extremes. There is a gradual shift as we move through the natural sciences, from physics through chemistry to biology and on to ecology, psychology and so on. The acceptance of axioms, and their use in constructing tight logical structures, become less significant, and large circles in argument become more significant. There is also a gradual shift in the sources of justification for the use of our methods of argument, whether axiomatic, circular or mixed, from the measurable empirical success of comprehensive theories, through the measurable empirical success of particular results, to the deepening of understanding. These two shifts bear a complex relationship to the balance between foundations and coherence that is captured by Susan Haack's idea of foundherentism (*Evidence and Inquiry*, chapter 4). As we move up the scale of disciplines, we do not simply move from the overwhelming importance of foundations in justification to the overwhelming importance of coherence. Axiomatic structures are a key source of coherence, and the success of theories in making sense of individual situations, a foundational type of justification, is important even in the least axiomatized disciplines. On the other hand, well-established fundamental principles of a natural science can provide support for particular beliefs within that science in a foundational way, and beliefs in the humanities can be tested by asking whether the overall picture that they give us makes sense, a coherentist criterion.

So we have justifications for many different types of belief. Even if no structured set of justifications can guarantee the correctness of a belief, that does not prevent structured sets of justifications from giving us the level of confidence in beliefs that we require in order to regard those beliefs as potential knowledge, so long as we do not expect certainty. Large circles of argument, and arguments from sets of axioms that have proved to be

fertile and successful, can confer plenty of confidence. They can certainly confer the rational account that Plato put into play as one way of distinguishing knowledge from true belief, although he ultimately rejected that option (*Theaetetus*, 201-210).

Circular arguments and contradictions

There is a further sceptical attack to consider. If we accept that belief in the conclusions of a circular argument can be regarded as justified by such an argument, then we run the risk of accepting that belief in a conclusion can be justified, even though belief in a contrary conclusion would have to be regarded as justified to the same extent. (Rival conclusions tend to be specific enough to be contraries rather than mere contradictories, but contradiction is still the result.) That would appear to be most unsatisfactory. We need to assess the level of this risk, and we need to consider how best to deal with it.

The risk is significant. There are plenty of examples in the humanities where two people start with the same data and, after conducting equally careful arguments, reach contrary conclusions. This can even happen in philosophy. (For some examples in another discipline, see Lamont (ed.), *Historical controversies and historians*.) We live with such outcomes. In contrast, if a contradiction arises in mathematics or in the natural sciences, particularly in physics or in sciences that are close to physics, we expect the experts to resolve the contradiction, or to advise us to suspend judgement pending a resolution. This reflects the greater significance in those disciplines of systematic structures that rest on axioms, or at least of fitting results into such structures, even if the results were first established by less systematic means. A systematic structure should be less vulnerable to contradictions than a set of large and overlapping circles which link propositions that are not structured in a clear hierarchy, so contradictions in the natural sciences are more likely than contradictions in the humanities to be regarded as symptoms of curable mistakes. Contrary conclusions in the humanities can be regarded as symptoms of mistakes, but even if they are, we may well see the mistakes as incurable, in the sense that it is not within our power decisively to resolve a conflict between two conclusions. Alternatively, but relatedly, we may see contradictions as the result of a

certain looseness in arguments in the humanities. An argument with a given starting point is not constrained by logic to move to only one possible conclusion. One specific reason why arguments in the humanities are not so constrained is that it is relatively easy to interpret evidence in a range of different ways. In the natural sciences, on the other hand, theories tend to come into collision with data in ways that make the collisions hard to mitigate by proposing alternative interpretations of the data. The price of re-interpretation can be a change in the foundations of the discipline. Given that those foundations are indeed foundational, in consequence of the hierarchical and tightly-knit logical structures that are the norm, such a move is only rarely plausible.

It would be tempting to respond to this sceptical attack by saying that as and when good arguments for a new conclusion that contradicted an existing conclusion came to light, that would simply take away the supposed justification for the earlier conclusion. That would be the right response in mathematics and in the natural sciences, but it would be too facile in the humanities. It is perfectly possible in the humanities for two people to adhere to contrary conclusions, and for an impartial observer to regard both of them as fully justified in holding their beliefs. The impartial observer might not regard either belief as knowledge. But the issue here is not that of whether beliefs do qualify as knowledge. It is that of whether they meet the standard of justification that is required in order for them to be candidates for the status of knowledge. Both beliefs could be candidates, without raising the spectre of the contradiction that would follow from a claim that they were both knowledge, and hence that they were both true. Furthermore, the disputants need not be required to suspend judgement by any principle that one should attach equal weight to the opinions of oneself and of one's epistemic peers. There is an argument that such a principle does not necessarily require the suspension of judgement in the face of disagreement on a contentious question. Disagreements on related questions can legitimately reduce the status in one's own eyes of those whom one would otherwise regard as peers (Elga, "Reflection and Disagreement", section 12).

The problem would not even be solved merely by saying that the appearance of an equally justified contrary belief should be regarded as negating the justification for an earlier belief. We must be able to handle

claims that beliefs are justified when equally justified contrary beliefs have not yet appeared, but when we know from experience of the discipline in question that they may very well appear. In mathematics, physics, chemistry and much of biology, this is not a great worry. Any existing justified belief might be overthrown, but the structures of the disciplines, and their high and well-defined standards of justification, make the risk low. In the humanities, we know that the risk is significant. So how can we ever claim that circularly justified beliefs in those disciplines are really justified, when we know that we are at significant risk of finding equally good justifications for contrary beliefs?

Everything hinges on whether we can maintain an extensional distinction between guaranteeing that a belief is true and conferring sufficient justification on that belief. The argument that the justification sceptic cannot establish global scepticism by using the purely structural form of his argument depends on our doing so. If the distinction can be maintained, then we can allow for some instances of sufficient justification for each member of a pair of contrary propositions, without requiring both propositions to be true and thereby generating a contradiction. Maintenance of the distinction will however only be sufficient to allow us to handle pairs of contrary propositions in this way, so long as they are proportionately rare. If we had to say, of a high proportion of the propositions that we asserted, that belief in a proposition and belief in a contrary proposition were both well-justified, we would not be able to regard our standards of justification as set high enough to be much of a guide to truth. Then they would not be appropriate standards to use in order to identify candidates for the status of knowledge.

The obvious argument against a distinction between guaranteeing that a belief is true and conferring enough justification on that belief, is that standards of justification would seem to be appropriate precisely when they would only be met by justifications for true beliefs. At the very least, the proportions of sufficiently justified beliefs in sets of beliefs should not often be much greater than the proportions of true beliefs in those sets. But we could meet that condition, even if sufficient justification did not guarantee truth. We should start at the axiomatic end of the scale of disciplines. If we can maintain the distinction there, we can be confident of being able to maintain it in the humanities too.

There are some disciplines, notably mathematics and some of the natural sciences, in which the criteria for regarding a belief as sufficiently justified to make it a candidate for the status of knowledge are so stringent that their proper use comes as close as we can expect to guaranteeing the truth of beliefs that satisfy the criteria. Despite this, we can still separate a guarantee of the truth of beliefs from their having sufficient justification. We can see this by seeing why justification without truth can be expected to be rare in those disciplines. We do not simply work from foundations, but work from foundations that are kept under review for what can be built on them. Foundations that allow a great deal to be built are favoured. Foundations that are unproductive, and foundations that lead to contradictions or to results that are at variance with observation, are discarded. That process of review bolsters the claim of beliefs that are based on the foundations that we use to be well-justified. It also increases the probability that those beliefs are true, because it incorporates into our methods a process that we can expect to help us to sort correct from incorrect foundations. In this way, we can give ourselves grounds to expect that if a belief has enough justification to be a candidate for the status of knowledge, it is also true. But we do not have to regard beliefs as sufficiently justified only when their truth is guaranteed. We can explain the link between justification and truth without incorporating a guarantee of truth into our criteria of sufficient justification, because we can see that the link is a consequence of the methods that we use. The relationship between use of those methods and justification is necessary, in that we regard beliefs in mathematics and in the natural sciences as justified when and only when they have been acquired using those methods, methods which in those disciplines include extensive requirements to test new results. In contrast, their use bears only a contingent relationship to the truth of our beliefs. Human fallibility limits the relationship to a contingent one, so we can have no expectation that justification and truth should be perfectly correlated. But a recognition of that contingency does not deprive us of an explanation of the link between justification and truth.

Having established that the distinction between guaranteeing the truth of a belief and its having sufficient justification can be sustained in the natural sciences, we may conclude that the distinction can also be sustained

in the humanities. The link between the use of certain methods and the truth of the beliefs that are acquired by using them remains contingent. The link between the use of those methods and justification is also contingent in the justification to method direction in the humanities, because we are more inclined than in the natural sciences to recognize the worth of results that are substantiated in unconventional ways. We can therefore claim to have defeated the sceptical argument that is based directly on the possibility of good justifications for contrary conclusions. But the sceptic would have one more line of attack.

This last line of attack would be to argue that the apparent scope to accept that two contrary beliefs could both have sufficient justification to make them candidates for the status of knowledge was simply a corollary of the absence from the humanities of anything that was definite enough to count as knowledge, aside from dates and other facts that could be catalogued and that were uninteresting in themselves. Loose definitions would be seen as the norm, alongside circular arguments. While the sceptic could take this line, it would involve a shift away from justification scepticism to something different. The claim would no longer be that for a belief to count as knowledge, it would have to be justified in a way that was not in fact possible. Instead it would be a claim that a belief could only count as knowledge if it had very definite content. Such a claim was most famously made by Karl Popper, with his criterion of falsifiability (*The Logic of Scientific Discovery*, chapter 1, section 6). But Popper's concern was with the natural sciences, rather than the humanities. It is open to the defender of the humanities to argue that they are not of the same nature as the natural sciences. They deepen our understanding in ways that do not depend on the making of strictly falsifiable claims. The study of history, literature, music and so on, and the study and practice of philosophy, are very effective in helping us to understand ourselves and our place in the world. That is a sign, although not a demonstration, that our work in the humanities is carried on in a rational way and that we should set some store by the conclusions that we reach. It is true that results in the natural sciences also deepen our understanding in ways that go beyond the factual content of those results. They can have a profound impact on our attitudes toward ourselves and toward the world. Darwinian evolution and quantum

indeterminacy are cases in point. But although one reason why we take seriously the broad implications of results in the natural sciences is that the natural sciences do make falsifiable claims that have been tested and that have not been falsified, it does not follow that only falsifiable claims can deepen our understanding.

Furthermore, we can make sense of the humanities and of their relationship to the natural sciences in terms of the scale to which I have referred, rising from the axiomatic, the logically structured and the precisely falsifiable claims of physics to the looser and occasionally contradictory claims of the humanities, where arguments in large circles are the norm. As we go up the scale, from physics to chemistry, biology, psychology, sociology and history, one approach gradually becomes proportionately more significant than the other.

I conclude that justification scepticism can be kept at bay, even though sceptics will never be silenced.

CHAPTER 6

Science

In this chapter, I discuss the sources of some questions in the philosophy of science, and then give a picture of our place as subjects in the objective world. In section 6.1, I outline the relationship between scientific results and their interpretation. Interpretation gives rise to questions in the philosophy of science, just as the exposure of propositions to propositional attitudes gives rise to epistemological questions. In section 6.2, I discuss the scope to avoid interpretation within part of physics. In section 6.3, I argue that interpretation enters into the practice of science in the higher sciences, so that it is inevitable. Our use of the concept of causation is a particularly striking example. In section 6.4, I explore some links between that concept and the concept of rational action. In section 6.5, I consider the debate between realism, anti-realism and structural realism. Finally, in section 6.6, I draw some threads together by discussing the relationship between the scientific conception of the world and our status as subjects within the world.

6.1 Science and philosophy

Science, like knowledge in general, is a fertile source of philosophical questions. As with knowledge in general, we can think of ways in which those questions might be avoided. To avoid epistemology, we would need to avoid bringing propositions into the picture, in any way that exposed them to propositional attitudes. We could not realistically manage that. The natural sciences give us a species of knowledge, and the questions of epistemology arise in relation to that knowledge. But other philosophical

questions arise too. To avoid those other questions, we would need to avoid any interpretation of our scientific results. That is, we would need to avoid applying concepts such as those of reality and of causation. In this chapter, I shall set out some reasons why that too would be beyond us, except in a very limited part of science.

We must first distinguish between the discovery and the contemplation of scientific results. My focus will be on the contemplation of results, and on the interpretation of them that is needed in order to allow that contemplation, rather than on their discovery. The distinction between discovery and contemplation is reminiscent of the distinction that is often drawn between the contexts of discovery and of justification. But it would not be appropriate to use that distinction, because my concern is not with the legitimacy of specific scientific results, but with some aspects of our broader relationship to our scientific knowledge.

The degree of interpretation that is required as part of our contemplation of scientific results depends on how much is expected of that contemplation. We therefore need to set a standard, in order to give the discussion definiteness. My standard will be that we must be able to contemplate results in a way that will allow us to make the next scientific advances. We must be able to make sense of what we know, in order to have an idea of what to try next. To that extent, contemplation feeds into discovery. The desire to understand the conditions of our being able to make the next advances limits the scope of the discussion here and in sections 6.2 to 6.5. I shall not be concerned with the general nature of the relationship between knowing subjects and the objects of their knowledge, a concern that would open up a whole range of philosophical concerns. I shall, however, discuss the relationship between ourselves and the world that we study in section 6.6.

We need clear notions of uninterpreted and of interpreted scientific results. Uninterpreted results are results that are not given in terms that are appropriate to our direct experience of the world. The direct experience that is meant here is experience of the world as one of objects that are real, that we perceive directly, that act causally on us and on each other, and on which we can act. Thus it is the experience that is available to perceiving agents in general, including zombies. Qualia are not the theme here. Where the terms that are appropriate to our direct experience happen to be the

scientifically fundamental terms, the results will be regarded as interpreted, even though there is no other formulation to be interpreted unless we go out of our way to create one. This may seem odd, but the point is that our use of terms that are appropriate to our direct experience is what gives a number of philosophical questions their purchase.

We can see how easily our statements of results come to be put in terms that are appropriate to our direct experience by considering something as simple as saying that rainfall causes a plant to grow. In saying that, we talk in terms of objects, their properties and causation. To hold back from doing so, we would have to limit ourselves to saying that the state of a whole system evolved in a certain way. In this example we would have a system of water molecules, of soil particles and of plant cells. Indeed, if we really wanted to drive causation out of the picture, we would be driven back to describing the whole system in terms of elementary particles, because any higher-level description would involve a view of entities, such as molecules and plant cells, that were seen as independently real, with their own properties, and as participating in causal chains.

The mathematical formulation of quantum mechanics would be one example of uninterpreted results, but so would any formulation of results, from any science, that was given purely in a formal language, where terms, possible statements using those terms and the logical relationships between those statements were defined, and a catalogue of statements that were accepted as correct was given, but no more was added. The point is that the formulation would only give the minimum that was needed to capture the logical structure and the current content of the science. (The content would be the set of statements that were accepted as correct. One must be wary of labelling them as true, because doing so might, depending on one's theory of truth, entail suppositions as to reality.) The formal language might be simple or complicated, and it might or might not allow for vagueness or modalities. The use of a formal language would not be necessary to ensure that only the minimum was included, but it would be the safest approach. The use of an informal, natural language would bring in associations that would be very likely to add extraneous elements to the uninterpreted formulation of the results in question, leading us to state those results in terms that were appropriate to our direct experience.

The use of a formal language to state results in an uninterpreted form would not remove the empirical element from science. We would say certain things using the symbols of the language, and we would refuse to say other things, in the light of the results of experiments. The things that we would say would correspond to the results that we regarded as correct, and the things that we would not say would correspond to the results that we regarded as incorrect or as not established. But even though there would still be a connection with experiments, and even though we might say that a given result summarized or predicted the outcome of a given experiment, we would not try to make sense of statements within the formal language by invoking such concepts as reality, laws of nature, objects and their properties, and causation. We would have no need to engage in that degree of interpretation merely in order to relate our formal statements to empirical data, so that we could use those data to sort the correct from the incorrect statements. We would recognize that there were constants and variables for individuals and for predicates in the formal language, which it would be natural to interpret in terms of real objects, properties and relations, and that there were universal statements, necessities and relations of implication, some of which it would be natural to interpret in terms of laws and causation, but we would not attach any significance to the mere possibility of adopting those interpretations. There are, however, two important limits to our freedom from interpretation. The first limit is that if we were to dispense with all interpretation in terms that were appropriate to our direct experience, we would place ourselves in a parlous state. Without some connections between the world as directly perceived and our words, our words would be bereft of meaning. The games in which we manipulated symbols would then be too idle to be worth playing, and it would become impossible to connect our games with empirical observations. An ability to dispense with the given type of interpretation of our scientific theories is parasitic on our practice, in everyday life, of linking words to worldly objects, to their properties and to their interactions. The second limit is that experimental practice implies suppositions as to reality and causation, a point to which I shall return below.

There is a connection between interpretation in the sense of imputing reality, objects, properties, laws and causation, and interpretation in the

sense of painting mental pictures or devising analogies that are related to our everyday experience. The latter type of interpretation implies the former type. If we delve into quantum mechanics, and support our understanding of the equations by seeing electrons as waves, then we take for granted the reality of those waves, just as we take for granted the reality of the ripples that we see when we drop a stone into a pond. If we understand the Earth's orbit by seeing the Sun as pulling the Earth toward itself, so that the Earth does not fly off at a tangent, then we see gravity as a real force that causes the Earth to change its course, just as we see the pull of a long rein as a real cause of a horse's trotting in circles round an exercise yard. Such visualizations bring with them the presuppositions as to objects, properties and causation that accompany any description in terms that are appropriate to our direct experience.

Despite this connection, interpretation in the sense of imputing reality, objects, properties, laws and causation is not the same thing as interpretation in the sense of painting mental pictures or devising analogies. If the latter were what mattered, the requirement for interpretation would be unduly dependent on contingencies of our mental capabilities, capabilities that varied greatly from one person to another, with many people at each level of capability. Some people are much better at thinking abstractly than others. The more skill and experience someone has within a given field, the less he will need visualizations and analogies. Another point to note is that the causation that we must avoid mentioning if we are to avoid the interpretation that leads us into philosophical questions is not merely the local causation that is most easily captured in visualizations. There is also the influence at a distance that is one of the obstacles to there being any adequate visualization of fundamental physics. If we equated interpretation with the painting of pictures, we would be likely to miss that point.

There is a further level of appreciation of scientific theories and results, beyond not needing visualizations and analogies. People could speak the language of mathematics, or some other formal language, as a native language that captured their experience perfectly. Statements in the formal language would amount to un-reformulated statements of direct experience. The concepts of reality, objects, properties, laws and causation,

with their usual senses, senses that reflect our everyday experience of the world, would not then be used when thinking about parts or aspects of the world that could be described in formal terms. That might be an ideal to which we should aspire. But most of us are not, and are not going to become, like that. If this special form of appreciation of statements in formal languages exists at all, it is rare. Its rare existence would not make the requirement for interpretation unduly contingent, in the sense of making that requirement vary widely across the population. The following arguments on the need for interpretation do not consider this possible escape from that need. They are arguments that we, as we are or as we might reasonably expect to become, mostly need interpretations of sorts that require us to face up to philosophical questions. They are not arguments that no conceivable rational being could dispense with such interpretations, or even that no actual human being could dispense with them. I shall also not concern myself with hypothetical people who would directly experience the world in a form that broke free of the usual ontology of objects that were clearly demarcated from one another and that acted on one another only locally. Such people would have the advantage that the form of their direct experience was more suited to an appreciation of the quantum world than was the form of most people's direct experience. But as we shall see, it is in the higher sciences that interpretation is in general impossible to avoid. That is where the talents of native speakers of mathematics, or of habitual users of non-standard ontologies, would, if they were widespread, affect the argument that follows. And it is not clear that those talents could be exercised at all in the higher sciences.

The point of avoiding interpretation

The appeal of uninterpreted formulations of scientific results is that if we were to confine ourselves to them throughout the sciences, many questions in the philosophy of science would fall away. Are hypothesized forces really there, or are they just artefacts of our theories? We would not care, because the results would come out the same either way. Do laws of nature follow from the properties of entities, or vice-versa, or both, or neither? Again, it would make no difference to the uninterpreted results. What is causation,

and in what sense do causes necessitate their effects? These questions would not arise, but we could still calculate that a system that was in a given state would evolve into another given state after a certain time. And so on. We use the concepts of reality, objects, properties, laws and causation in interpreting our scientific results, not in giving their uninterpreted formulations. Even if we use existential quantifiers in stating our results, that does not in itself import the concept of reality. The quantifiers are merely symbols in a formal game until we interpret them. (This is subject to the point made above, that we must not let our games become too idle.) Alternatively, we could stop at the interpretation of our scientific results, and not go on to reflect on the significance of our interpretations. That would also save us from philosophy. But that would be a wilful refusal to philosophize, not the discovery of a principled reason not to pose philosophical questions.

Uninterpreted formulations would also have the advantage that they would help to confer the objectivity that is rightly craved by science. We do not want a science that reflects our personal, cultural or even species-specific inclinations. Our choice of areas to investigate will inevitably reflect those inclinations, but whatever results we happen to obtain should not reflect them. Uninterpreted formulations hold out precisely that prospect. The use of a de-humanized language, the language of mathematics or of logic, would be ideal for this purpose. Aliens might use different mathematical and logical concepts, but if we ever met them, their and our mathematical and logical languages would rapidly merge. What better proof could there be of the de-humanization of our science than an ability to share it with non-human rational beings, and to do so with no loss or change of content?

The question is, could we manage this? Could we contemplate the results of science in a way that avoided interpretation, and that therefore allowed us to avoid many of the traditional questions in the philosophy of science, while we kept enough of a grip on the content of the results to allow us to make the next scientific advances? It depends on the type of science. It turns out that we could do so, some of the time, within a narrow field that I shall call equation physics, but that it would become both harder and less important to do so, the further we moved away from that field.

6.2 Equation physics

Some physicists prefer to avoid interpretation of their theories altogether. They restrict themselves to generating more equations and solutions to equations. Anyone who has tried to picture the results of quantum mechanics will understand why. We cannot construct pictures that are entirely faithful to the mathematics. Moving on to the sense of interpretation with which I am concerned, we cannot even impute reality and causation in their ordinary senses. Phenomena such as nonlocality mean that our everyday understanding of reality and of causation would just get in the way. If we try to interpret the results of quantum mechanics in ways that are completely faithful to the theory, we both fail and give ourselves headaches. We also run the risk of drawing conclusions from our interpretations that do not really follow from the mathematics. So we had better stick to the mathematics. Mathematics itself involves concepts, but they are precisely defined within the theories that use them. That conceptualization within a mathematical theory is to be distinguished from the interpretation with which I am concerned, interpretation in terms that step outside the theory.

The advantage to the working physicist of setting out results in an uninterpreted form is that it allows him to get on with his job and to communicate his results to other physicists, so that further progress can be made. We can have confidence in the results, and we can use them to derive other results and to make useful devices. If physicists would not go beyond what they could interpret in terms that were appropriate to our direct experience, they would make very slow progress. If they based further conclusions on such interpretations, they would make a great many mistakes.

That much makes an adequate case for avoiding interpretation. The question is, can we do so? We largely can, to the extent that a mathematical formulation says all that need be said in order to capture the full content of the physics in question, while still being graspable by suitably trained human beings in a way that allows them to make the next advances. I shall refer to the part of physics in which a mathematical formulation is both comprehensive and comprehensible as equation physics. I introduce this

term in order not to bring in the assumptions about its extension that would come with the use of an established term, such as "fundamental physics" or "mathematical physics". Equation physics is the part of physics in relation to which we can adopt Heinrich Hertz's view of Maxwell's theory of electromagnetism, "Die Maxwell'sche Theorie ist das System der Maxwell'schen Gleichungen", "The Maxwellian theory is the system of the Maxwellian equations" (*Untersuchungen über die Ausbreitung der elektrischen Kraft*, page 23).

There are two provisos to the claim that within equation physics, interpretation can be avoided. The first proviso is that there is one circumstance in which even results within equation physics definitely cannot be contemplated, to a standard that allows the next advances to be made, without leading us into interpretation. This is when in order to make the next advances, we need to design experiments in which we see ourselves as acting on entities, or see some entities or events that involve them as causing events to occur. I shall return to this circumstance in section 6.5, in connection with the debate between realism and anti-realism. The second proviso is that even outside the context of such experimentation, physicists who want to make progress may need to think in causal terms. Christopher Hitchcock has pointed out that much depends on the state of the science. He notes that references to causation occur often enough in scientific papers, even in fundamental physics, and that only in phases of maturity of a science, when most of the ground that it covers has been systematized mathematically, may it be feasible to discard the language of causation ("What Russell Got Right", section 3.7). Furthermore, such phases of maturity may be followed by fresh phases of immaturity, as physicists introduce new theories or tackle new phenomena. Despite these provisos, there are times when results within substantial parts of equation physics can be contemplated, and the physics can be advanced, without the benefit of interpretation in terms that are appropriate to our direct experience. Even in a science's mature phases, there is still work to be done to advance the science, filling in the gaps and exploring tensions that may eventually reveal the need for radical re-formulations that will plunge the science back into immaturity. I shall use the term "mature equation physics" to identify the part of equation physics that may be regarded at a given time as mature.

There may be no time at which the whole of equation physics is mature, but there will be times at which some parts are both mature and open to being advanced. We should, however, note that not all philosophical questions can be dismissed, even in relation to mature equation physics. One question that cannot be avoided on principled grounds, even when we only reflect on results within mature equation physics and do so without interpreting their content, is the question of the general character of laws of nature.

The boundaries of equation physics are hazy. In some parts of the natural sciences, we nearly always need interpretation in terms of reality, objects, properties, laws and causation in order to have enough of a grip on our results to be able to advance to the next results, or even in order to be able to see difficulties with our current theories. Interpretation even enters into the formulation of the theories themselves, as will emerge in section 6.3. But there is no clearly-defined point at which interpretation starts to be required to any significant extent. We can, however, safely include some current theories of fundamental forces within the scope of equation physics. Richard Healey gives a thorough analysis of the state of play in *Gauging What's Real*. It becomes clear to a reader of the book that the business of interpreting theories of fundamental forces is separate from the business of formulating and using those theories. Engaging in interpretation, Healey asks whether or not gauge theories push us toward particular ontological commitments. His conclusion is that gauge theories do not carry significant ontological implications (ibid., chapters 4 and 8), and indeed that we have reason to be wary of any such claims that might be made on their behalf (ibid., chapter 9). Healey does, however, acknowledge that gauge theories have significant metaphysical implications, ones that may astonish us when we reflect on the contrast with our everyday experience (ibid., section 4.5). Two general points may be noted. The first general point is that the possibility of avoiding ontological commitments is not directly relevant to the thesis that there is a part of physics that can be practised without interpretation, because the question of ontological commitments only arises once we decide to turn our minds to interpretation. But Healey's conclusions are still reassuring, in that they would limit the extent to which my thesis would need to be revised if it should turn out to be impossible to avoid turning our minds to

interpretation. The second general point is that Healey makes it possible to argue for some of his conclusions by adopting a particular type of mathematical representation of gauge theories, loop representations. This indicates that the availability of philosophical results in this area can depend on the contingencies of our mathematical progress. Such dependence need not disturb us. It is, after all, only the achievement of a certain level of mathematical sophistication in our formulation of physics that allows us to identify equation physics as a distinctive part of natural science at all.

We can also include in equation physics the classical mechanics of macroscopic regions of space-time and of the macroscopic entities therein, both in relativistic and in non-relativistic forms. Fortunately it will not matter, for the subsequent discussion, precisely how much of physics falls within equation physics. All that matters is that there are some parts of physics that are fundamental, that are the basis for a great deal of science, whether or not a full reduction of science to those parts of physics is even theoretically possible, and that do, in at least some of their phases, fall within the scope of mature equation physics. (It might appear that classical mechanics, at least when non-relativistic, should count as interpreted, as defined in section 6.1. The fundamental formulation might appear to be in terms that were appropriate to our direct experience, with objects being seen as moving around and as being pushed and pulled by forces. But the more sophisticated Hamiltonian formulation is given in abstract mathematical terms.)

The nature of equation physics makes it easy to see how questions in the philosophy of science, as opposed to more general epistemological questions, might come to be dismissed. If the mathematical description of a system says all that need be said, many philosophical questions have no purchase because the concepts on which those questions rely are creatures of interpretation in terms that are appropriate to our direct experience. They are not natural outgrowths of equation physics. There is no concept of reality, or of causation, in the equations themselves. We cannot even separate out objects and properties. There are simply equations that describe a system. Those equations include a variable that is commonly given the name “time”. Changes in the values of other variables that are correlated with changes in that variable correspond to the evolution of the

system over time. One striking piece of evidence for the avoidability of interpretation is that we have now been doing quantum mechanics for nearly a century without being able to decide, to everyone's satisfaction, whether it is about the world or about our knowledge of the world (Goldstein, "Quantum Theory Without Observers – Part 1", page 42; Allori and Zanghi, "What is Bohmian Mechanics"). In the higher sciences, it is much harder to avoid using the concepts that give philosophical questions their purchase.

6.3 The higher sciences

It might seem at first glance that we could work with uninterpreted formulations of scientific results in sciences other than equation physics. We could define particular types of molecule, or species of animal, and work with our definitions and the data that were gathered from experiments, while excluding any interpretation that went beyond the definitions and logical constructs that were based on those definitions. This would amount to the extension beyond equation physics of the project of Carnap, using Ramsey sentences and developments of them (Carnap, "Testability and Meaning"; Carnap, *Philosophical Foundations of Physics*, chapter 26). There would be no desire to understand the significance of our theories, merely a desire to get to grips with the observations and to improve our ability to do so by improving our theories. That limited ambition would enable us to avoid at least some of the philosophical difficulties that surround Carnap's project. In particular, we would not have to worry about the reality of entities, even though Ramsey sentences do implicitly refer to entities, by description, because we would abstain from the interpretation of our theories. We would likewise be undisturbed by the central concern that is set out in Cei and French, "Looking for structure in all the wrong places", that Carnap's project would only give us structuralism and would not give us realism.

Alas, such an approach would leave us unable to make the next scientific advances. The use of uninterpreted results would become harder, the further we moved away from physics in general and from equation

physics in particular, both because it would become progressively harder to give definitions of the relevant entities that were both precise and all-encompassing, and because statements of results in purely formal terms would become less and less comprehensible. There are few things to say about an elementary particle, and they can all be said in precise, mathematical terms. There are rather more things to say about a molecule, and vastly more things to say about an animal. The uninterpreted statement of results in a fully formalized theory might remain theoretically possible in sciences that were distant from physics, but it would be completely impractical. We need to speak more loosely when we practise the higher sciences. That leads us to use terms that are appropriate to our direct experience. It might be thought that we could stick to a loose, incomplete, formal expression of a science, and steer clear of interpretation, but that would not allow us to make the next advances. We need comprehensible accounts that are full-bodied and definite, even at the cost of their being incomplete and of their being given in terms that use broad-brush concepts. We need such accounts both in order to formulate the next scientific questions, and in order to comprehend the material that we already have at our disposal and that we can use to help us to answer those questions. We can get what we need by using the concepts that are appropriate to our direct experience. We structure our information by talking about objects, such as molecules or animals, and their properties. That gives us full-bodied and definite pictures of the world, even though we are well aware both that we do not know all of the properties of objects, and that the properties that we do identify are multiply realizable at the micro-physical level. We also use the concept of causation to tell comprehensible stories about what happens. We notice when such stories are not fully satisfying because not all significant causes have been identified, or not all of the obvious potential interrupters of causal processes have been ruled out. What happens as we move up the scale of sciences is that concepts that are, in equation physics in its better moments, used only to interpret the science and not to formulate it, come to enter into the formulation of the science. The boundary between formulation and interpretation is broken down, and the interpretation that leads to philosophical questions becomes unavoidable.

It is at this point appropriate to consider the implications of adopting the ground for regarding interpretation as unavoidable that was mentioned in section 6.1. Interpretation is to be regarded as unavoidable in the practice of a given science when it is needed in order for us to be able to make the next advances in that science. The adoption of this criterion might seem to limit the significance of the need for interpretation. The use of interpretation might seem to be a mere matter of convenience, of no philosophical significance. Once the next round of advances had been made, interpretations could be discarded. But the philosophical significance of interpretation is not to be minimized in this way. I identify the need for interpretation in order to identify a route that leads to questions in the philosophy of science. Since we have a strong urge to advance our scientific knowledge, we are led to those questions. The further question of what status we should accord to the concepts that we have to use, such as the concepts of reality, objects, properties, laws and causation, is to be answered within the philosophy of science, once the practice of that philosophy has been pressed on us. The fact that we were led to the philosophy of science only by the practicalities of the scientific enterprise would not in itself mean that we should assign to those concepts the low status of mere crutches for our limited minds.

Fortunately, the formulation of results in terms that are appropriate to our direct experience becomes both easier and less distortionary as we move away from physics. Animals, for example, are objects of our direct experience. We directly perceive many of the differences between them that lead us to classify them in different species. Moving back down the scale of sciences to chemistry, we directly perceive that immersion in hot water will cause a sugar lump to break up, and that two given elements regularly react with one another in constant proportions. (At least, we can perceive that, so long as the elements are supplied in visible forms and quantities.) When thinking about molecules, we can at least sometimes picture atoms as plastic balls with holes in them, and the bonds between them as wooden struts that fit into the holes, without leading ourselves into contradiction or doing too much violence to the scientific theory. The fact that only limited violence would be done makes it less important than it is in equation physics to avoid using terms that are appropriate to our direct experience. A significant

diminution in the violence that is done to theories does not, however, necessarily require a substantial movement up the scale of sciences. The Newtonian mechanics of macroscopic objects can be stated in terms that are appropriate to our direct experience, without doing violence to at least some respectable mathematical formulations of the theory. There is also no suggestion that the concepts that are appropriate to our direct experience are adequate to state all of the results, anywhere on the scale of sciences. All of the natural sciences include bodies of knowledge that are too sophisticated to be expressed in such terms, and that can only be stated mathematically. It is not that statements using the concepts that were appropriate to our direct experience would be distortionary. There could not be any such statements of the sophisticated results.

We cannot conclude from this alone that as we move away from physics, there is a sharp diminution in the danger of drawing mistaken conclusions from expressions of results in terms that are appropriate to our direct experience. But at least the growing ease with which the conclusions can be understood should increase the ease with which mistakes can be detected. Eventually scientific disciplines shade into the humanities, where definitions that are both precise and adequate to their subject matter are neither expected nor feasible, arguments in large circles take over from systematic deductions from basic principles, expressions of results in terms that are appropriate to our direct experience are the norm, and those expressions are adequate to the subject matter because the subject matter is itself the social world in which we live and act, the world that we routinely describe in such terms so as to conduct our lives. What historian, for example, would venture to define the concept of progress or the concept of class in a way that would allow his definitions to be used in the construction of arguments with precise logical structures, with any reasonable expectation that historians generally would agree that his definitions were either acceptable or useful? Some historians do make such attempts, but the results, while they may be philosophically significant, are not impressive as works of history. Hegel and Marx both ignored Aristotle's injunction to adapt their expectations of exactness, of certainty and of the universality of conclusions to their subject matter (*Nicomachean Ethics*, book 1, section 3, 1094b11-27).

I shall now illustrate the significance of the distinction between equation physics and the rest of science by reference to two topics, causation and realism.

6.4 Causation

The concept of causation is fundamental to the conduct of most of the natural sciences, even though it need not always be used in the mere contemplation of a formal characterization of results. It is easy to make causation disappear when contemplating results within mature equation physics, but it becomes harder, and eventually impossible, as we move up the scale of sciences. We cannot really avoid saying that the introduction of a catalyst causes a chemical reaction to speed up, or that a change in temperature causes a seed to germinate. Causation is also a concept that is appropriate to our direct experience. And if we cannot avoid using the concept of causation, then we must face a range of philosophical questions that relate to the nature of causation and the reality of the entities that participate in causal events. We might avoid those questions so long as we only used the concept of causation, and never mentioned it, but as soon as we came to think about our scientific knowledge, we would be faced with the questions. And not thinking about the concept of causation would amount to an unprincipled refusal to engage with philosophical questions, given that we had to use the concept.

There might or might not theoretically be some reductive account of the results of the higher sciences that would take us back to a system of elementary particles, the initial state and the evolution of which we could describe mathematically without using the concept of causation. We would need such a reductive account because the substantial entities that are the subject matter of chemistry and of biology cannot be fully characterized using precise definitions that capture every physical detail, so long as we confine ourselves to the terms of those sciences. A system that can only be described in one way in the terms of a higher science can be realized in many different ways at the micro-physical level. But even if a reductive account were theoretically possible, it would be wholly

impractical. A reductive science of the objects that were the subject matter of any one of our higher sciences would be quite different from that higher science in its unreduced form. We need to work with selective and micro-physically incomplete characterizations of the entities with which most of our sciences deal, and we need to use the concept of causation in order to stitch together the members of a temporal sequence of selective descriptions of the world. The resulting narratives themselves only mention a small selection of the causal relationships that could be mentioned. The sense of selection that is meant here is not that of taking a small proportion of the micro-physical facts, but that of selecting a few of the many different higher-level facts that could be abstracted from the entire micro-physical base. Incompleteness is therefore not the direct correlate of selection, in the way that it would be the direct correlate if selection meant that some micro-physical facts were selected and that others were ignored. Rather, the incompleteness consists in the fact that multiple realizability would make it impossible to determine all of the micro-physical details, if one started with a narrative that was given in higher-level terms.

The link between selective descriptions and our use of the concept of causation is no accident. The concept of causation is ideally suited to the task of arranging a sequence of selective descriptions in a coherent and significant narrative, because it is a concept that we have developed in order to state our direct experience. We have naturally developed for that purpose concepts that help us to assemble narratives that report only the significant facts. We say that one event causes another, as when we say that the impact of a porcelain cup on a concrete floor causes the cup to shatter. In so doing, we focus on particular features of the objects involved, the brittleness of porcelain, the speed of the cup just before it hits the floor and the hardness of the floor, in order to explain an outcome that is itself characterized in a specific way, as the shattering of the cup, rather than by giving the positions of all of the small pieces of porcelain around the room. (One interesting view of the concept of causation that would be perfectly consonant with its role in allowing the construction of coherent and significant, but highly selective, narratives, while not being implied by the existence of that role, is the view that the validity of causal judgements is not absolute, but that

it is relative to the theories and situations in which the judgements occur. See Menzies, "Causation in Context".)

Causation and rational action

I shall now set out two connections between the concept of causation and the concept of rational action. The first connection is that the practice of rational action justifies the selection of facts and the use of selected facts in constructing causal narratives. The second connection is that the practice of rational action justifies our use of the concept of causal necessity. The concept of rational action that is involved here is the concept that was discussed in chapter 4, that of action that we regard as rational because it has been chosen by a process that makes appropriate use of the evidence. To be more specific than in chapter 4, I am here concerned with actions that are both directed toward particular goals, and reasonably likely to lead to the achievement of those goals.

There is no novelty in making connections between the notions of action and of causation. The scope for a link was identified by Thomas Reid, in the words, "The conception of an efficient cause may very probably be derived from the experience we have had in very early life of our own power to produce certain effects" ("Essays on the Active Powers of Man", essay 1, chapter 5, page 524). A different line of thought on the connection between action and causation is represented by G. H. von Wright's paper "On the Logic and Epistemology of the Causal Relation", which cites the role of deliberate manipulation in identifying causal relationships and in distinguishing causes from effects. That line can be developed into an argument that agent causation is conceptually prior to event causation, in that one must have a grasp of the former concept in order to have a grasp of the latter concept (Lowe, *Personal Agency*, section 6.8). Several connections between our concepts of causation and of agency are explored in Ahmed, "Agency and Causation". But the focus of most of the work of this nature has been on the analysis of causation and on how, as a matter of psychology, we get to grips with the notion of causation. My concern is rather to find justifications in our practice of rational action for our practice of relying on highly selective accounts of the world, and for our assumption that there is causal necessity to be found in the world.

Rational action, selection and causal narratives

When we observe a cause and its effect, we observe a sequence. We may observe a sequence many times, but we have no guarantee that the sequence will always be the same. This is not to say that we are at all likely to be mistaken in predicting effects in straightforward cases, such as dropping porcelain cups onto concrete floors. If an expected effect did not follow, we would rightly look for an explanation, rather than shrug our shoulders and say that regularities were never guaranteed to continue. There is nothing wrong with formulating laws of nature and expecting them to apply in ordinary instances of causation. To return to the example of the cup, laws of the mechanics of macroscopic objects allow us to calculate the force of the impact, while laws of chemistry and of materials science explain how molecules in a piece of porcelain are linked, and why the application of certain forces will be followed by fractures. But this picture is incomplete. We need to look at the ways in which laws of nature are linked with one another.

The most basic laws are given at the level where the concept of causation need not be used, in equation physics. Less basic laws, of chemistry or of materials science, can be built up from those basic laws, but they are not straightforward laws that apply universally. Rather, an instance that conforms to such a law represents the confluence of applications of several more basic laws, together with the absence of an indefinite range of countervailing conditions, whether states or events, that might have prevented the usual outcome. Thus the process of combining ingredients to make porcelain will lead to the formation of bonds that give porcelain its usual strength and brittleness, but only so long as the proportions of the ingredients are right, there are no impurities of types that would interfere with the formation of the bonds, the firing temperature is within a certain range, and so on. Then we have laws such as the law that a porcelain cup that is dropped onto concrete from a height of one metre will shatter. Such laws represent a further stage in the combination of laws and the ruling out of countervailing conditions, although the main possible countervailing conditions may cover a narrower range, and may be more obvious to the lay person, than at the previous stage where chemistry and materials science came into play. It is when we move away from the basic level of equation

physics, at which the concept of causation need not be used, that we have to start formulating regularities that cannot be formulated as exceptionless laws. This happens because the complexity of the objects that are involved, and the complexity of the sets of basic laws that bear on an instance of a regularity, both give wide scope for countervailing conditions to interfere. The countervailing conditions that would need to be ruled out cannot all be listed to allow exceptionless laws to be formulated, because their range is indefinitely wide. This lack of exceptionless laws is a symptom of the fact that the concepts that are used in the higher sciences are not appropriate to the task of capturing every detail of the world. The concepts that are used in equation physics are appropriate to that task. As we move up to the higher sciences, the concepts that are used have to capture complex interactions of the underlying mathematical laws, and they must do so in a broad-brush way if the laws that use the higher-level concepts are not to be hopelessly convoluted. The notion of "broad-brush" is given content by defining it in terms of the multiple realizability, at the micro-physical level, of a state of the world that is characterized in a single way when using a given set of broad-brush concepts. The scope for unusual interactions of the underlying laws, interactions that would be triggered by details that were not captured by the broad-brush concepts, creates the scope for exceptions to higher-level laws. The triggering details would be among those of the countervailing conditions that would be hardest to foresee.

The account that has just been given may be compared with one that is given by James Woodward ("Causation with a Human Face", section 4.5). Woodward is, like me, concerned with the consequences of the coarse-grained nature of the variables that feature in the higher sciences, and with the scope there, but not in fundamental physics, for interference in causal processes by extraneous factors that are not ruled out by descriptions of systems in coarse-grained terms. It does not follow that the theories that are enunciated in the higher sciences are hopelessly unstable. As Woodward points out, coarse-grained specifications, both of causes and of their effects, allow us to establish relationships between causes and effects that are more stable than the relationships that would obtain between incomplete fine-grained specifications (ibid., pages 88-89). Detailed differences in the state of the world at an earlier time, differences that

would only show up in a fine-grained description, may lead to differences in the state of the world at a later time that would create exceptions to any tidy laws that related incomplete fine-grained descriptions of the world at earlier and later times. Coarse-grained descriptions of the world at earlier and later times may obscure the detailed differences in such a way that laws that relate coarse-grained descriptions are not particularly vulnerable to exceptions that would be generated by such differences. Another interesting point of comparison is the view that the occasional failure of laws in the higher sciences should be attributed to the occasional failure of the combinatorial laws that determine the resultant of a number of individual forces (Rupert, "Ceteris Paribus Laws, Component Forces, and the Nature of Special-Science Properties").

My remarks above show that it is not coincidental that the move away from science that can be given wholly in terms of equations, and the move away from exceptionless laws, occur together. Rising complexity drives both moves. That complexity would make any putative equation science at higher levels impractical to use in taking the next steps in the enlargement of our knowledge. It would also make exceptionless laws in the higher sciences impractically convoluted. The picture that is painted here is, however, still one of causation as it features in causal explanations that conform to the covering-law model. There is no move to singular causal explanation, in the form in which it is seen as falling outside that model (Woodward, "A Theory of Singular Causal Explanation", section 3).

Faced with the complexity of the everyday world, we must rely on the selection of facts, and on the combination of the selected facts into coherent and significant narratives that use the concept of causation. We make sense of what happens to a particular cup by giving a highly selective narrative, in which its being dropped and the consequent impact cause it to shatter, but from which most of the facts are omitted because their absence does not detract from the narrative. We simply say that one event, the impact, is the cause of another event, the shattering. Our selection of facts is based on our interests. Variations in which factors matter to us? Are we, for example, interested in changing the way in which we make cups, so as to reduce the frequency of breakages? Or do we like porcelain sufficiently to insist on continuing to use it, but wish to do something to avoid

breakages? We have here a species of contrastive focus (Dretske, "Contrastive Statements"). We may contrast porcelain cups with melamine cups. Alternatively, we may contrast the dropping of porcelain cups onto concrete with the dropping of porcelain cups onto carpet. We focus on one type of fact, the composition of the cup or the nature of the floor. The practical outcome may be that we change the way in which we make cups, or that we resolve only to take porcelain cups into carpeted rooms.

We need something to justify the practice of selecting some facts and not others. Here is the role for our practice of rational action. Our desires to achieve certain results motivate our selections of some facts rather than others: we concentrate on what matters. But we can analyse the process in a little more detail, and thereby see how our success in rational action justifies our practice, in our scientific enterprise, of selecting facts and of using the concept of causation to stitch together selective descriptions of the world in order to create narratives.

When we choose actions in order to achieve certain results, we ignore many details. If someone prepares a budget for a business in order to facilitate its management during the year, she will not pay any attention to the type of spreadsheet program that she uses to do the work. Such details are irrelevant because they only affect some equally irrelevant details of the final result, such as the type of computer file in which the finished budget is recorded. When we have a goal, we describe the goal in a way that brings out what actually matters. We do not list all of the changes that we might effect by achieving the goal. We then give equally sketchy descriptions of the actions that we must perform in order to achieve the goal. Nothing else would be practical. All of this is perfectly everyday. The significant points are first, that it is the practicalities of getting things done that drive us to be selective, to distinguish between what matters and what does not, and second, that we still generally achieve the desired results. Another essential tool in our getting things done is an understanding of causal chains. We need to know what will happen if we initiate certain changes. We need to identify the right causal narratives, the ones that take account of the important facts and disregard the irrelevant ones. If we want to remove ice from some machinery, it is useful to know that a local rise in temperature will cause the ice to melt. It is also useful to know that some

details are irrelevant, so that we do not waste time making choices in relation to those details. We do not need to fuss over which source of heat to use, because any source that will not damage the machinery will do.

Our experience of rational action draws our attention to the fact that we can use limited information in order to bring about desired states of affairs that are described in limited ways. The significance of this obvious fact is that it shows us that selective descriptions can still be perfectly good descriptions. We take risks by relying on selective descriptions. There may on occasion be some unnoticed factors which mean that our usual ways of achieving certain results will fail. But we know from experience that the risk is low, and that we can get a perfectly good grasp of the ways in which things work without burdening ourselves with too much information. (A related point is that we can often get away with using fast and frugal heuristics, which rely for their frequent but not universal success on the fact that some cues are consistently useful indicators of the correct answers to questions. See Gigerenzer, *Adaptive Thinking*, chapter 8, and especially pages 175-177.) To put the point in mathematical terms, we learn that the world is very often not chaotic, in the sense that ignoring some of the available information does not place us at much risk of getting results of our actions that are very different from what we intended. That discovery justifies the conduct of sciences above the level of equation physics on the basis of highly selective descriptions of the world, descriptions that are then stitched together to give causal accounts in which only selected causal relationships are mentioned. The descriptions and the accounts lack precision, but that does not leave them devoid of accuracy.

We could equally well justify our presumption of the frequently non-chaotic nature of the world by pointing to the fact that we are able to construct elegant scientific theories that allow us to make a wide variety of successful predictions. We should also note a certain interdependence between the lack of chaos and our ways of selecting facts. If we approach the world in the ways that we do, we find that we can achieve things. If we were to select facts in different ways, we might find that we could not achieve much at all. That is, the world may only come across as non-chaotic if we approach it in the right ways, if we find the joints at which to carve (Plato, *Phaedrus*, 265e). Our practice of rational action shows not only

that there are right ways to be found, but that we have found some of them. Finally, there is an anthropic argument that there must be some ways of approaching the world that will make it come across as non-chaotic. If there were no such ways, creatures of our sophistication would not have evolved, and could not have survived for long even if they had somehow come into being. But that argument, unlike our experience of rational action or the content of our successful scientific theories, would give no indication of what the right ways to approach the world might be.

Rational action and causal necessity

There is another role for our practice of rational action in relation to causation. The practice can justify our use of the concept of causal necessity, or of causal raising of the probabilities of given outcomes. (I shall refer simply to necessitation. I wish to leave room for probabilistic causation, but not to enter into the debate as to whether it exists.) The justification is based on the fact that the notion of acting in order to achieve a certain result makes no sense if we do not expect the world to have to respond to our actions. There is another apparent justification, based on the fact that we decide to act, which I shall examine because its limitations reveal the benefits of being aware of when subject origination and the boundary concept of the subject are implicitly invoked.

We have the concept of a chain of events, the strength of the links in which goes beyond mere constant conjunction, even constant conjunction that will prevail throughout the life of the Universe. The relevant sense of justification is that of showing us that we are right to conclude that instances of causal necessitation reflect the nature of the world. As with the use of selective accounts, there are alternative justifications which do not rely on our practice of rational action, and I shall set out one such alternative. But it is vital that we should have some justification for our use of the concept of causal necessity. A concept of causation that did not incorporate the idea that causes necessitated given outcomes would not be our concept of causation, as it is used in the conduct of our sciences and in our daily lives. The challenge to which we must respond was set by David Hume, when he claimed that all that we ever saw was constant conjunction (*An Enquiry Concerning Human Understanding*, section 7, part 2).

The justification that is based on our practice of rational action is that there is a logical link between rational action and causal necessity. We act in order to achieve specified results. It only makes sense to perform an action in order to achieve a given result if we see the action, or some more basic action that is involved in it, as bringing about the result, except when the action is itself a basic action and the desired result is merely that the action should have been performed. (A basic action is an action that the agent performs directly, without having to do anything else.) Whenever an action stands in a causal chain and there are subsequent links in the chain that lead up to the desired result, we see the action as causing the desired result. An action does not make the desired result inevitable, because things can go wrong. But the action must be seen as pushing the world toward the desired result. Rather than seeing the concept of causal necessity as based on observations of constant conjunction, we can see it as a depersonalized abstraction from our practice of action in order to change the world. Then rather than be concerned, as Hume was, that we do not observe the causal necessity that we believe to exist, we can see our perception of the necessity as an outgrowth of our awareness that we must assume necessity in order to make sense of the concept of rational action. The concept of rational action is one that we understand not in the abstract, but on the basis of our lived experience as agents. As Goethe had Faust say, "Im Anfang war die Tat", "In the beginning was the deed" (*Faust*, part 1, "Studierzimmer", line 1,237). We should, however, note that the concept of causal necessity that is supported by this argument is defined by the argument in terms of pushing the world. That is enough to capture our everyday sense of causal necessity, but a more refined concept of causal necessity would need a more sophisticated support. The argument that is given here leaves the field open to a wide variety of different philosophical accounts of causation. What this argument shows is that if we want to make sense of rational action, we must philosophize about a concept of causation that includes causal necessity.

Much would need to be added in order to complete the picture. I shall here remark on only one piece of the jigsaw, although it is a large piece. We would need to have senses of ourselves and of our actions before we could use our understanding of our practice of rational action to justify our

use of the concept of causal necessity. Fortunately, there is reason to think that this would not be a stumbling block. We could adopt the approach that is used by Lucy O'Brien. She gives a fundamental role to the agent's awareness of her actions, claiming that this awareness is an awareness by means of the production of actions rather than by means of the reception of data, and that this awareness does not presuppose a capacity for first-personal reference and thought (*Self-Knowing Agents*, page 88). O'Brien's approach could supply the piece in the jigsaw that I would need because the agent's awareness of her actions does not depend on a prior grasp of the concept of causation, so we would not be faced with circularity. The reason why it does not so depend is that it includes awareness of basic actions. Basic actions can, like non-basic actions, be followed by changes in the external world. We can go on to identify those changes as effects, casting our basic actions in the role of causes. (For our knowledge of basic actions, and the significance to O'Brien of the category of basic actions, see *Self-Knowing Agents*, pages 160-168.)

Another apparent justification for our use of the concept of causal necessity that is based on our practice of rational action is worth examination, but largely to reveal its limitations. The starting point is that we deliberately initiate rational actions. As Wittgenstein remarked, "Voluntary movement is marked by the absence of surprise" (*Philosophical Investigations*, part 1, section 628, page 137). Going beyond the absence of surprise, the agent decides whether, as well as when, to act. There are then effects in the external world. The world responds to something that it did not know was coming, even though the agent knew. So it does not seem that the constant conjunction between an action of a certain type and an effect of a certain type can be merely a constant conjunction between events of the given types within the world. In order to make sense of the whole picture of deliberate action by the agent that startles the world into a response, we must suppose a stronger bond than that of temporal coincidence within the world. That stronger bond can be described as causal necessity.

This justification for our use of the concept of causal necessity can, however, only show that its use gives us a picture of deliberate action that broadly makes sense. If we bring all elements of the picture into sharp

focus, we come to see unresolved difficulties. Specifically, the argument relies on our seeing a deliberate action and an external effect. The action and the effect are regarded as not apt to be seen as two members of a pair of events within the world that, for all we can tell, merely coincide. But they can only be regarded in the required way if we do not embed the action, on all sides, within the world's network of causes and effects. The picture is implicitly one of the subject under the boundary concept, engaging in subject origination from a place that is not within the world, and thereby taking the whole world by surprise. But if that is the picture, then we do not have a picture of one event within the world causally necessitating a second event within the world. The second event is seen as within the world, but the first one is not. So we do not properly justify our use of the concept of intra-world causal necessity. There are delicate questions here as to whether a movement of the agent's body should be counted as an action which is seen as not embedded in the world, or as an effect within the world of a separately identifiable volition, but wherever we drew any dividing line, the main point of the argument would stand. Something would have to be seen as not embedded in the world, if it was to be seen as taking the whole world by surprise.

I can now fulfil the promise that I made in section 2.4 to give a reason why we cannot be sure that our concept of causation would be shared by all rational beings, and hence why we cannot be sure that the content of our natural sciences would be accessible to all rational beings. Necessity is central to our concept of causation. A concept of causation that did not incorporate some form of necessity would not be adequate to the conduct of our sciences. We need to see natural laws as having the force of law, entitling us to demand reasons for failures. It is possible that as a matter of psychology, rather than of logic, we need to reflect on our practice of rational action in order to get a grip on the concept of causal necessity. We cannot be sure that all rational beings would conceptualize their lives in terms of rational action, nor can we be sure that they would have alternative routes to the concept of causal necessity. The risk of their being cut off from a concept of causation that incorporated some form of necessity may strike us as small, partly because an understanding of the world that did not use such a concept would be one that would strike us as

very strange, if indeed we could comprehend it at all. But the risk must be noted.

I shall now turn to a justification for our use of the concept of causal necessity that is independent of rational action, and that is preferable to the extent that it works. Its starting point is the obvious first stage in a response to Hume's challenge. This is to remark that we can easily do better than take at face value the conjunctions that are visible in everyday life. We can explain those conjunctions in terms of finer details and underlying laws of nature. Those details and laws may in turn be explained by even finer details and even deeper laws. Thus the fact that an impact with a certain force will cause a porcelain cup to shatter is explained by the arrangement of the molecules, and by the laws that dictate how they are held together. The molecules can then be analysed into atoms and the bonds between them, while the laws that dictate how the molecules are bound to one another can be explained by deeper physical laws that govern the behaviour of electrons. Eventually we reach the fundamental level at which causation can drop out of the picture and there are only the equations that describe the evolution of systems, but most of the way down, we need to use the concept of causation in order to give a comprehensible account.

The possibility of getting down to the level at which causation need not be seen holds out one prospect of bypassing rational action in the search for a justification for our use of the concept of causal necessity. Suppose that we could, in theory, always get down to the level of equation physics, even though we might not yet know how to do so. Suppose also that we could regard the equations that governed the evolution of systems, as described at that level, as brute facts that explained necessity at higher levels but that were themselves a source of necessity that we could accept without any explanation of how they managed to be such a source. Then we would not need anything more to justify our use of the concept of causal necessity. (The presumption of the necessity of the relationships between basic laws and higher laws would need to be justified, but that would not be a causal necessity. It would be more in the nature of a conceptual necessity.) The claim that all instances of causation could be appropriately related to processes that were the subject matter of equation physics is open to challenge. But it is worth noting that some challenges to

the ambitions of what might look like an impossible reductionist project would not be in point, because all we would need would be a correspondence between each token of causation at a high level and tokens of low-level processes that ran in accordance with equations. Thus the frequent lack of type-type correspondences would not be a valid objection. It might in any case be possible to transmit necessity upward from equation physics, even if it was not possible to reduce the higher sciences downward.

An account that was based on brute necessity at the level of equation physics would have one great advantage over an account that was based on our practice of rational action. It could explain the fact that causal necessity existed, or if one preferred a different idiom, the fact that the world was amenable to description by propositions, some of which started with the word "necessarily". An account that is based on our practices can justify our use of the concept of causal necessity, but it cannot explain necessity itself. To the extent that there is such necessity, it is independent of our practices. It must be, because if it exists, it existed long before there were any rational beings. It would be implausible to see it as created by our thoughts, given that the evolution of our brains depended on the regular behaviour of entities in accordance with natural laws. Having said all that, it is still worth considering the way in which our practice of rational action can justify our use of the concept of causal necessity. The generation of necessity at the level of equation physics, or its transmission to higher levels, might be found to be unsatisfactory for some reason.

6.5 Realism

I shall now turn to the debate between realists and anti-realists. Structural realists have joined the party in recent decades, although structural realism is heir to a philosophical tradition that stretches back a century or more. Both realism and anti-realism encompass a wide variety of more specific positions, but we can characterize them in broad terms. Realism is the view that the unobservable entities that are posited by our theories really exist in some sense, and anti-realism is the view that they should be regarded as mere artefacts of our theories. I shall here identify two sources of the debate.

The first, and indirect, source is that the concept of rational action requires us to bring propositional content into the picture, because we see ourselves as acting on the basis of reasons. The propositions that state our reasons are in turn exposed to propositional attitudes. Once we bring propositions into the picture in a way that exposes them to propositional attitudes, we can go beyond a desire to know what is the case and ask about the significance of our claimed knowledge. We are rapidly, although not inevitably, led to the question of whether our knowledge states how the world is (realism), or whether it merely gives an empirically adequate account of the world (anti-realism). The notion of propositional content moves us up a level and leads us out of the network of theory-laden scientific concepts within which Bas van Fraassen argues that we should stay. He argues that if we do stay within that network, that will keep us from taking a quizzical look at our science as a whole, thereby ensuring that we avoid awkward questions about realism (*The Scientific Image*, pages 80-83). But once we have a way out of the network, it is entirely appropriate to take that way and see where it leads.

The second, and more direct, source of the debate is the inevitability of our use of the concepts of causation, in the higher sciences and in immature equation physics, and of action, in our life generally. Our use of these concepts pushes us toward a realist view. The push might be resisted on a careful examination of the connections between concepts, but the push is still there. When we then reflect on realism, we find that it is not an unproblematic doctrine.

The push toward realism that is supplied by the concept of action has been clearly stated by Ian Hacking, in his famous remark about electrons and positrons being fired at a ball of niobium, "If you can spray them then they are real" (*Representing and Intervening*, page 23). It is very hard to take the things on which we act, or on which we rely in our plans of action, to be anything other than real, whether or not they are observable. Action is distinguished from fantasy by the fact that it involves our changing a real and often recalcitrant world. This does not make a knock-down case for realism. It might be that our concept of action should be changed. But any attempt to refute the case for realism would require us to enter into philosophical debate.

It is true that results, in any science, can be contemplated without reference to the concept of action, and can be so contemplated in a form that at least sometimes allows the next advances to be made. But the concept of action has a way in, and can push us into the philosophical debate, when the next advances must be made by carrying out experiments in which we see ourselves as acting on entities that we see as referred to, whether implicitly or explicitly, by our existing results. In the higher sciences, entities tend to be referred to explicitly anyway, and we are already faced with the question of realism for that reason. Experimental practice just adds to the pressure. In equation physics, the concept of action is not supported in its task of confronting us with the question by other explicit references to entities. But it can still do the job. The type of contemplation of existing results that is implied by the design of experiments exposes even the practice of mature equation physics to the philosophical question of realism. This does not, however, undermine the general claim that mature equation physics has an interesting immunity to many philosophical questions, including the question of realism. Mature equation physics can often, although not always, be driven forward by theoreticians who stay within the mathematics and who do not have any immediate need for new experiments to be designed. The immunity is there, even though it is imperfect.

The concept of causation also supplies a push toward realism. Unlike the push that is supplied by the concept of action, this push exists even when we merely contemplate results in any science in which we use the concept of causation, even without thinking about new experiments. We only expect real things, or events that involve real things, to cause other things to change. Hegel has a nice line, "Was wirklich ist, kann wirken", "What is real, can have effects" (*Wissenschaft der Logik*, part 1, book 2, section 3, chapter 2.B). We have a strong urge to turn this round and say, "Was wirkt, ist wirklich", leading us to take it that whatever we correctly see as having causal powers is real. (The qualification of correctness is important. It is not meant to put in question the existence of causal powers in general, but it is meant to block an inference to the reality of imagined causes.) Again, it might be possible to resist the push toward realism. But again, it could only be resisted by entering into philosophical debate.

Thus the question of realism is pressed on us by our use of the concepts of action and of causation. We are pushed toward a broadly realist view. We may wish to resist, either because we believe that there are good arguments against realism, or because we consider that the push is only a push toward a pre-philosophical type of realism that should be replaced by a more sophisticated doctrine. But does our use of these two concepts also lead us to a particular answer, even after the debate has been conducted? The philosophical question is not that easy to answer. There are complex considerations on both sides. But the attractions of realism, as highlighted by the words of Hacking and the adaptation of Hegel that were given above, are hard to ignore. It is also noteworthy that a leading attempt to finesse the issue, Arthur Fine's introduction of his natural ontological attitude, has been forcefully argued to be a form of realism in disguise (Fine, "The Natural Ontological Attitude"; Musgrave, "NOA's Ark – Fine for Realism").

We should also consider structural realism. Whether it could satisfy our natural hunger for something real would depend on what would need to be seen as real in order to satisfy the hunger. We would need to consider ontic structural realism. Epistemic structural realism leaves open the question of what exists, so it would in itself neither feed us nor starve us. Ontic structural realism comes in many varieties, from an eliminativist version in which it is argued that there are no individuals, only relational structures, through versions in which individuals have only relational properties, or in which all irreducible properties of individuals are relational, to versions in which the existence and the identity of individuals depend on the relationships between individuals. It would be a considerable step for most of us to come to be satisfied with the primary reality of structures and relations, even if individuals were taken to have some derived reality, but it would be a much smaller step for people who were already familiar with contemporary physics.

Finally, it is not surprising that the debate over realism should have the origins that are indicated here. The debate can be avoided within mature equation physics, where we have no need to interpret the equations or to see any causation. At least, it can be avoided if we leave to one side the indirect source of the debate in propositional attitudes, and also refrain from

designing experiments. If we wish to interpret our equation physics, that is a different matter, and we are rapidly led into the debate over realism. Beyond equation physics, the boundary between theory and interpretation breaks down, leading us to make claims of reality within our theories. We also need to use the concept of causation, both in the higher sciences and in immature equation physics. Then the debate over realism cannot be avoided on principled grounds, even by the practitioners of the sciences concerned, although they may just assume realism in order to put the debate on one side and concentrate on obtaining more results. One interpretation of structural realism is that it is an attempt to exploit a concentration on structures rather than on things in order to extend across science the benefits of being in the safe zone of the practice of equation physics, where the debate need not often rage. Specifically, structural realists tend to identify structures that are given in mathematical terms. This was so in John Worrall's choice of the example of theories of light ("Structural Realism: The Best of Both Worlds?"), and the tradition has been continued through structural realists' tendency to focus on physics. It has been argued that structural realism in the form that is advocated by Worrall is only suited to the mathematical sciences (Newman, "Ramsey Sentence Realism as an Answer to the Pessimistic Meta-Induction", pages 1,377-1,378). But there are also arguments that structural realism can be applied to other sciences (Ladyman et al., *Every Thing Must Go*, chapters 4, 5 and 6).

6.6 The subject in the objective world

In this section, I shall return to the themes of chapters 2 and 3. I shall identify a role for the boundary concept of the subject in making sense of the relationship between the subject and the world as it is regarded when we practise the natural sciences.

The scientific ideal is a view from nowhere, to borrow Thomas Nagel's title, although to be precise it is not a view but a conception, because all views are from somewhere. This conception of the world is detached in the sense that it keeps the observer, as observer, out of the

picture. He is not assigned a location, and the act of observation is not mentioned when describing the world. This notion of a detached conception of the world must be distinguished from the notion of an unbiased view of the world, a view of how the world is that is uninfluenced by our theories, even though the two notions are often discussed together. Hilary Putnam, for example, writes of the God's Eye View (*Reason, Truth and History*, chapter 3). As Putnam makes clear, the assumption that there is such an unbiased view is the assumption that there is one true way that the world is, carved up into objects and their properties in a certain way. But Putnam also discusses the notion of a view from above the world, looking down on, for example, lots of brains in vats (ibid., pages 49-52). That latter notion is the notion of a detached conception, so long as we do not read "from above the world" as assigning a location to the observer. Putnam's particular concern with the notion that we might all be brains in vats creates particular connections between the notion of a detached conception and the notion of an unbiased view, but in general, the two notions can and should be kept separate. The English language does not help. The two notions correspond to two senses of the word "objective".

The difficulty with taking the observer out of the picture is that we are aware that it is we who are the observers. We are subjects in the objective world, and we should be able to connect the idea of our being subjects with the notion of a detached conception of the world.

When we consider specific observations, there is no great difficulty. We can both be, and see ourselves as being, in one part of the world while studying something that is in another part, and the fact that we are in the same world as the thing that is studied does not matter. We can distance ourselves from the thing that is studied in a way that allows us to take ourselves out of the picture. A geologist can study a mountain, and a biochemist can study molecules, without any relationship between the observer and the observed featuring in statements of the results, even though the geologist may be standing on the mountain and the molecules may be inside the biochemist's own bloodstream. There is plainly no problem with taking the observer out of the picture here. At the same time, there is no loss of contact with our lives as subjects. We can do science, directly apprehend the objects of study and understand our apprehension

of those objects, all at the same time. Even if some objects are too small or too fleeting for us to see them unaided, so that we have to use special instruments, we enter into a relationship with the objects that can be grasped as a whole. We and the objects are both within space-time. We can imagine both ourselves and the objects on that broad canvas, while still being able to describe objects without mentioning our relationship to them. As a useful bonus, a scientific account of that broad canvas, an account that includes ourselves, our instruments and the objects of study, makes perfect sense of our experience so far as it can be captured on the basis of the naturalistic conception. If we see ourselves, our instruments and what we observe together, we can fully understand the origins of the activity in our brains that corresponds to our experience. The originator conception still has its special role in doing justice to the experience of making free but controlled choices, but the scientific account leaves room for that.

When we move on to our relationship to the world as a whole, serious difficulties emerge. The concern is not with particular experiences, but with the framework that makes experience possible. There are two possible orders of conceptual priority. We can start with the subject and work outward to the spatio-temporal world. Alternatively we can start with the world, insert people into it and then establish their status as subjects. The first option leaves us with two different relationships to the world. We see subjects or we see the world in detail, but not both at the same time. I shall explore this option first, before arguing that the second option can do the job. The second option is intuitively more satisfactory than the first to anyone who is innocent of philosophy. It aligns conceptual and historical priority. This is an advantage if we seek a philosophy that will make us feel at home in the world.

The first option is given to us in Kant's *Critique of Pure Reason*. His view of the spatio-temporal world goes hand in hand with his view of space and time themselves. For him, space and time were pre-conditions of all of our science, the very forms under which we apprehended the empirical world. We could study space and time, but their fundamental role meant that we were in great danger of falling into paradox. The risk of error when thinking about space and time is demonstrated by the first antinomy, in which a proof that the world had a beginning in time and has spatial limits

is set alongside a proof that the world neither had a beginning in time nor has spatial limits (*Critique of Pure Reason*, A426-433/B454-461). In Kant's view, those who fall into such paradoxes do so because they do not properly understand the relationships between the subject who thinks and perceives, the world as it is in itself and the world as we perceive it. His solution, transcendental idealism, is set out briefly at A490-497/B518-B525, although the full picture is given only by the *Critique of Pure Reason* as a whole. One component is that the true thinking subject is not in the phenomenal, spatio-temporal world. Another component is that we can, indeed must, take everything in the world as perceived by us to be real, even though things as perceived are not as they are in themselves. (I have no desire to come down on one side or the other of the debate between those who interpret the Kantian system as involving two worlds and those who interpret it as involving one world. For a survey of the debate from a one-world point of view, see Allais, "Kant's One World: Interpreting 'Transcendental Idealism' ".)

The Kantian system is a complete system, in which the subject keeps in touch with the world. But the Kantian philosophy does not give us a single relationship to the world. There are two relationships, that of the philosopher and that of the perceiving subject. When we act as self-conscious philosophers who both express and contemplate the whole system, we have a conception of the subject, with his access to phenomena, although we can say nothing about the subject. When someone acts as a perceiving subject, he does not, as a subject, appear in his own picture of the world at all, although he can observe his own body. This whole approach might end in disaster, a possibility that I shall note here but shall not pursue because my aim is to give the approach a fair wind, and then argue that we can move beyond such a two-relationship approach. The risk of disaster has been set out by Adrian Moore, who argues that it is impossible for a Kantian coherently to state his conclusion that "we cannot know anything or intelligibly say anything except from the point of view of possible human experience" (*Points of View*, page 126). Moore can, however, allow the conclusion to reflect limits on ourselves that can be shown, even though they cannot be stated, and he can identify the utterance of the Kantian conclusion as the outcome of an inevitably doomed attempt

to state those limits, an attempt to eff the ineffable (ibid., chapter 7). An alternative way forward would be to claim that when we acted as philosophers and contemplated the whole system, we could become aware of the fact that there were limits to what we could know, limits that reflected the bounds of sense (both linguistic sense and sense-perception), but that even as philosophers we could not see both sides of those limits, and could not say what the limits were. Such an approach would not vanquish the threat to the Kantian system that Moore identifies, but it would confine the threat somewhat.

The nature of the two-relationship approach is brought out most clearly by considering our relationship to space and time. We cannot explore space and time while we remain in the philosophical relationship to the world. We can set out their place in the system, as forms that allow us to experience objects phenomenally, but in order to explore them, we need to transfer to the empirical relationship and become subjects who perceive objects. We then suspend our philosophizing and become true scientists. Indeed, Kant stated in his 1770 *Inaugural Dissertation* that geometrical properties of space, such as its three-dimensionality and the existence of only one straight line between two points, could not be derived from a universal concept of space but could only be discerned concretely, in space itself (Ak.2: 402-403). He made the same point in the *Critique of Pure Reason* when he stated that in order to do geometry, we need not just concepts, but intuitions (A46-49/B64-66).

Any exploration of space and time that could be carried out on the basis of the philosophical relationship to the world would be limited to the identification of properties of space and time that Kant's overall theory required. Such an argument could not get far. At best, we could only argue for a few topological properties. The leading candidate for a required property would be that time should have only one dimension (compare *Critique of Pure Reason*, A31/B47). Kant's own forays into this territory did not establish much in the way of specific properties. In his 1747 *Gedanken von der wahren Schätzung der lebendigen Kräfte*, he gave scientific rather then philosophical reasons why our space had three dimensions, but acknowledged that spaces with different numbers of dimensions were possible (Ak.1: 24-25). In his 1768 *Von dem ersten Grunde*

des Unterschiedes der Gegenden im Raume, he argued for the existence of absolute space on the basis of chirality, but even with this help from intuitions of left and right hands, he did not argue for specific properties of space. In the *Inaugural Dissertation*, he remarked on the continuity of time (Ak.2: 399-400). We certainly could not go so far as to argue for a particular geometry of space. Kant's overall theory, abstracted from the particular intuitions that we in fact have, would not require space to be Euclidean. Having noted these limits on the conduct of geometry without the benefit of intuitions, we should also note that in Kant's view, given that space did have whatever geometry it in fact had, that geometry was inevitable. We could never find any objects that broke the laws of geometry, because those laws gave the forms of all of our experience (*Inaugural Dissertation*, Ak2: 404; *Critique of Pure Reason*, A22-25/B37-40). But even the intuitions that we in fact have need not drive us to a Euclidean geometry. We could recognize that our intuitions were consistent with a geometry that was Euclidean locally, but non-Euclidean on a larger scale. It so happens that in the century after Kant, his noun "Mannigfaltigkeit", "manifold", came to denote a space that was Euclidean locally but that was not necessarily Euclidean globally. One must however add both that this possibility was not available for Kant himself to consider, and that if he had been able to consider it, he would have been faced with the challenging notion of a non-Euclidean geometry that was to be taken seriously as a characteristic of the empirical world, but that was not reflected in our intuitions because they were all local rather than global.

The Kantian system allows us a very limited appreciation of space and time in our capacity as philosophers who grasp the human subject as well as space and time, and a much fuller appreciation of space and time in our capacity as human subjects who experience objects phenomenally. (The limitation to human subjects was, for Kant, significant. He claimed that we could not judge whether other thinking beings would be bound by the same conditions as those that limited our own intuition: *Critique of Pure Reason*, A27/B43.) The Kantian starting point is the problem of how to make sense of the world as it appears to us and of our relationship to that world. The starting point is not a world that we construct in a theoretical way, without reference to any subject, a world under a detached

conception. If we start there, difficulties initially fall away. It is not hard to start there, but we then have to retrieve the subject. This is the second possible order of conceptual priority that was mentioned above.

We can first construct a mathematical model of space-time and then locate objects, including human beings considered merely as objects, within that space-time. This gives us the world under a detached conception. Our mode of access to the world as thus conceived is theoretical, not perceptual. That will make it hard to retrieve the subject, but in the meantime it allows us to dismiss the four antinomies (*Critique of Pure Reason*, A420-461/B448-489). These puzzles arose because Kant wanted to make sense of the world in terms that were appropriate to our experience, even though he discussed things that were beyond direct experience, such as the beginning of the world and the microscopic level of detail at which we might see ultimate particles. The first antinomy is the clearest case. The proof that time had a beginning relies on the argument that if it did not have a beginning, we would have taken an infinite time to get to the present, and that a completed infinity is impossible. But it is perfectly possible mathematically for us to be located somewhere in particular on a line that stretches away endlessly behind us. The proof that space is finite relies on our inability to intuit the construction of infinite space from finite parts, save in a disallowed infinite time. But if we take a mathematical model of infinite space, and assume that this space exists independently of our insertion into it and of our perceptions, we have no need to intuit its construction. We simply give the mathematical formulation that defines it. The proofs that time and space are infinite rely on the assumption that if they were not, they would have to sit within empty time and empty space. But there are perfectly good mathematical models of finite but unbounded space-time, which do not imply that it sits within some larger space-time that is otherwise empty. There are other ways to see off the first antinomy. Peter Strawson gives some arguments (*The Bounds of Sense*, pages 176-183). But while Kant's specific arguments are vulnerable to detailed criticism, only Kant's own solution, or the alternative of first preparing the spatio-temporal world, then inserting people, can protect us from the full range of possible arguments in Kant's style, arguments that start from the subject and work outward to the world. Similar things can be said about the

other antinomies. The second antinomy plays on our everyday understanding of the notions of breaking down solid entities into parts, and of constructing entities out of components. But we can give mathematical formulations that capture infinite divisibility, and formulations that capture the lack of it, without any reference to such intuitions. The third antinomy plays on our everyday understanding of natural causation and of freedom, but in a theoretical model of the world, the former would feature as an idiom in which we could express regularities in sequences of events, whereas the question of the latter would not arise. The fourth antinomy plays on notions of the necessity of the world and of things within it. It would not have any purchase in relation to a world that was modelled theoretically. That world would simply be there.

We must now insert subjects into the world as modelled. The easy stage is to insert human beings. If we do that, we can still have theoretical rather than perceptual access to the world. We can work out what the human beings could perceive and what they could do, given their locations and their powers, and we can work out the corresponding perturbations in their grey matter. Although we cannot in practice have a science of human beings without using terms that are appropriate to our direct experience, that practicality is not relevant to the central point, the contrast between theoretical and perceptual access to the world. We can in the current context use the terms that we also use to describe our direct experience, without tainting the theoretical nature of the access to the world that is posited here. Terms can be borrowed for new purposes without bringing the implications of their original uses in their train. The practicality matters more when we come to work out a route by which we might have reached our current understanding of the world, and wish to avoid circularity in that route. I return to that concern below.

The difficulty comes after we insert human beings. We need to make this populated world our own, a place in which we live as subjects. We need to create a satisfactory link between the world as conceived theoretically and the world as experienced by ourselves, not just as experienced by human beings who have been inserted into the world. There are three separate challenges here. The first challenge is to explain how someone realizes that she is herself a given human being within the world as

conceived theoretically, someone who perceives that world and who acts on it from a given location, and to explain how she aligns egocentric space and objective space. I touched on this challenge in section 4.3, and shall not discuss it further here, but shall rely on the rich literature that already exists and that rightly places stress on the links between perception, action and embodiment. The second challenge is to explain how a human being, seen in the world as conceived theoretically and therefore seen as fully embedded in the causal network of the world, can also be seen as a free subject. The third challenge is to show that the easily-banished antinomies do not return to exact their revenge, once we place ourselves as subjects in the objective world.

The second challenge is the one that the boundary concept addresses. If someone brings herself under that concept, she sees herself as a point of origin, and therefore as free in her deliberations, as well as rational. Kant would not have been able to use this approach. One reason why he could not locate the intelligible subject in a pre-existing spatio-temporal world was that the subject would then have been caught in the web of causation, making freedom inconceivable. Actual freedom rather than the perception of freedom was required, if we were to be potentially moral creatures. But if we can make do with the perception of freedom, as has been argued here, although it has been argued in relation to our self-conception generally rather than in relation to our status as moral creatures, and if we can tolerate the degree of mystery that follows from the lack of any account of how people might manage to be points of origin, then we can locate ourselves as free subjects within the spatio-temporal world. Use of the boundary concept allows us to see ourselves, from the outside, as free subjects, as centres of deliberation, choice and action who are not mere links in causal chains. It gives each person a way to apply the status of a subject to herself when she contemplates space-time with herself, as a human animal, within it. The sense of being a boundary of the world, rather than being within the world, is vital here. That sense prevents the dissolution of the subject into a set of interacting components when we look at the world as one big mechanism. It makes our status as subjects secure. We can accord ourselves a substantial subjective presence, rather than the non-presence of the Kantian intelligible subject. The benefit of

that substantial presence can be felt not only in relation to our status as free subjects, but also in relation to our status as rational subjects. This is so even though that latter status can be sustained by models of our mental processes such as the ones that have been proposed by Enç, Frankfurt, Frankish and Bratman, and that were discussed in section 1.5, without recourse to the boundary concept.

We can proceed in this way, without risk of circularity, because the concept of the subject is not needed in order to state the most basic scientific theories, the theories of equation physics, even though the concept of the observer is needed to state some, but not all, interpretations of quantum mechanics. The basic scientific theories are enough to give us space-time and matter. We can therefore conceive the world according to science, at least in the basic terms of equation physics, terms that do not rely on the concept of causation and hence are not at any risk of reliance on our sense of rational action. That is enough to give us a world in which we can place human beings, although initially without any detailed description of them because we do not yet have the full range of our sciences. We can then draw on our sense of action, ascend to self-consciousness, see ourselves as engaging in rational action and bring ourselves under the boundary concept. We then have enough to align egocentric space and objective space. We also then definitely have access to the concept of causation, so that we can build up the full range of our sciences and give detailed descriptions of human beings. (If the concept of rational action is not needed in order to give us access to the concept of causation, we can have our sciences at an earlier stage.) Our sciences allow us to refine the alignment of egocentric space and objective space, as we come to understand why it is that certain actions will lead both to certain changes in the world, and to certain changes in our perceptions and in our scope for further action. This account is not put forward as actual history, but as a rational reconstruction in order to show that the final position could legitimately be reached from the starting point.

The location of ourselves as subjects in the world may look like a verbal conjuring trick, even the apparent content of which is obscure, never mind the obscurity of what is really going on. In fact, there is no trick because there is so little to do. We are simply aware of ourselves as subjects

in the world, and continue to regard ourselves as subjects when we think theoretically about the world with ourselves in it. Identification of the role of the boundary concept does not change anything. It merely uncovers a way in which we can regard ourselves as subjects in the independent world.

The third challenge remains. The antinomies melt into air once we change our starting point from the subject to the pre-existing world that we access theoretically. If we then insert ourselves as subjects, rather than merely having a theoretical conception of human beings with their powers of sense and of action, do the antinomies return?

They do not return. I shall not analyse Kant's arguments in detail, but shall indicate why there are no problems in the areas that he identifies. The first antinomy can be dismissed both because if space-time is taken to be primary, there is no need to intuit its construction, and because there is no clash with our intuitions in the idea of a space-time in which our journeys neither come up against some limit nor take us indefinitely far from where we start. Making the theoretical conception of the world primary allows us to dismiss the second antinomy, because the question of whether or not objects are infinitely divisible is to be answered by reference to physics. Faced with what is currently regarded as an ultimate particle, we can visualize it as a ball and ask whether or not it could be cut in two. But that would be an interpretation of the science in terms that were appropriate to our direct experience, and it would distort the science. Physical theories tell us that our intuitions should not always be used when discussing the nature of the world. The third antinomy concerns natural causation and freedom. Use of the boundary concept allows us to reconcile a supposition of freedom with the fact that only natural causation exists. The fourth antinomy draws attention to the fact that we want to explain the existence of the world as a whole, but that all of our attempts to do so come to grief. That is a disturbing truth about the limits of our knowledge. But all of the explanations that we are ever likely to have will be scientific, and they are only likely to give us relative necessity. We have to live with that fact. Starting with theoretical access to the world may make it easier to accept this limit on our powers than it would be if we started with perceptual access. It certainly does not make it harder. It also changes the form of the question from that in which Kant posed it. We are not tempted to give

answers in the style either of his thesis or of his antithesis. Instead we transfer the question to cosmology, where it is both more tractable in detail and more obviously unanswerable in principle.

If we use the boundary concept in the way that is proposed here, adding a status but not adding any natural property to the human being as an entity in the world as conceived theoretically, then we can both recognize our status as subjects and recognize that this status depends entirely on physical entities within the world, on our bodies and above all on our brains. The structure and the spatial unity of each body and of each brain explain the unity of consciousness, at a time and over time, which is an essential part of our self-conception but which has sometimes puzzled philosophers. That unity is primarily a philosophical puzzle, rather than the scientific puzzle that is to be solved by noting the physical unity of each human body and working out the operation of the extensive connections that exist within each human brain, but not between brains. There are significant questions about how bits and pieces of neural activity all over a single brain are brought together in a unified consciousness, but it looks as though it should be possible to answer those questions by scientific means (Crick and Koch, "A Neurobiological Framework for Consciousness"; Singer, "Large-Scale Temporal Coordination of Cortical Activity as a Prerequisite for Conscious Experience"; McFadden, "Synchronous Firing and Its Influence on the Brain's Electromagnetic Field"). If we can take it that such questions will eventually be answered in satisfactory detail, and if we can see ourselves as subjects in the world, rather than merely as objects in the world and as subjects in some strange other sense, then there is no separate philosophical puzzle of the unity of consciousness to solve. We can see ourselves as subjects in the world by using the boundary concept because that gives us a special view of the physical objects that we are. It does not give us things that are additional to those physical objects, or additional properties. That restraint keeps the subject as philosophers see it close enough to the animal as natural scientists see it for philosophers to be able to rely on the consequences of the physical unity of the animal. In the rare cases in which injuries or malfunctions lead to a loss of unity of consciousness, the physical reality and the concept of the subject stay in line. There is indeed a loss of unity,

which may be either total or partial, both at the physical level and at the formal level that interests philosophers, although the loss of physical unity may reflect the loss of some co-ordinating function, rather than anything so obvious as a physical break in a neural pathway. Our usual concept of the subject must be modified for people with such injuries or malfunctions. Susan Hurley uses rare brain structures to argue that there is no necessary connection between neuroanatomical unity and the unity of consciousness (*Consciousness in Action*, essay 5). But while the connection may not be necessary, it is the one that in fact obtains in nearly all people. The analyses of consciousness, and the explanations of its unity or disunity, that are needed when neuroanatomical unity is lacking are special approaches that are needed for special cases.

Armed with the boundary concept of the subject, we can start with a mathematical model of space-time, then insert ourselves into the space-time that we have modelled and instantly feel at home. That insertion of ourselves as subjects into a pre-existing space-time allows us to see ourselves as never having been fettered by the bonds that David Hume identified with the words, "Let us chase our imagination to the heavens, or to the utmost limits of the Universe; we never really advance a step beyond ourselves" (*A Treatise of Human Nature*, book 1, part 2, section 6, page 67). Furthermore, we can apply the originator conception and see ourselves as deliberating with mastery, and hence as free and self-directed in the exercise of our rationality. That goes along with application of the boundary concept. In short, the boundary concept and the originator conception of ourselves allow us to find a place in the world that is our place, a place that does justice to our humanity.

Bibliography

No editions are specified for classic works that are available in many editions, except when references in the text have been given in forms that can only be used with specific editions. The online and the printed editions of specific issues of journals occasionally have dates of publication that are one month apart.

Ahmed, Arif. "Agency and Causation". Chapter 6 of Huw Price and Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford, Clarendon Press, 2007.

Aleksander, Igor. *The World in My Mind, My Mind in the World*. Exeter, Imprint Academic, 2005.

Allais, Lucy. "Kant's One World: Interpreting 'Transcendental Idealism' ". *British Journal for the History of Philosophy*, volume 12, issue 4, November 2004, pages 655-684.

Allori, Valia, and Nino Zanghi. "What is Bohmian Mechanics". *International Journal of Theoretical Physics*, volume 43, numbers 7-8, August 2004, pages 1,743-1,755.

Alston, William P. *Beyond "Justification": Dimensions of Epistemic Evaluation*. Ithaca and London, Cornell University Press, 2005.

Anscombe, G. E. M. *Intention*. Oxford, Blackwell, 1957.

Aristotle. *Nicomachean Ethics*.

Bains, Sunny. *Physical computation and embodied artificial intelligence*. PhD thesis, Open University, 2005.

Baker, Lynne Rudder. *Persons and Bodies: A Constitution View*. Cambridge, Cambridge University Press, 2000.

Barnes, Barry, and David Bloor. "Relativism, Rationalism and the

Sociology of Knowledge". Chapter 2 of Martin Hollis and Steven Lukes (eds.), *Rationality and Relativism*. Oxford, Blackwell, 1982.

Bekenstein, Jacob D. "Information in the Holographic Universe". *Scientific American*, volume 289, number 2, August 2003, pages 58-65.

Bergmann, Michael. "Reidian Externalism". Chapter 3 of Vincent F. Hendricks and Duncan Pritchard (eds.), *New Waves in Epistemology*. Basingstoke, Palgrave Macmillan, 2008.

Berlin, Isaiah. "Two Concepts of Liberty". Essay 6 of Berlin, *The Proper Study of Mankind*, edited by Henry Hardy and Roger Hausheer. London, Pimlico, 1998.

Blackburn, Simon. *Ruling Passions: A Theory of Practical Reasoning*. Oxford, Clarendon Press, 1998.

Boghossian, Paul A. *Fear of Knowledge: Against Relativism and Constructivism*. Oxford, Clarendon Press, 2006.

Brandom, Robert B. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA and London, Harvard University Press, 1994.

Bransen, Jan, and Stefaan E. Cuypers (eds.). *Human Action, Deliberation and Causation*. Dordrecht, Boston and London, Kluwer Academic Publishers, 1998.

Bratman, Michael E. *Structures of Agency*. Oxford, Oxford University Press, 2007.

Bruner, Jerome. "Life as Narrative". *Social Research*, volume 71, number 3, fall 2004, pages 691-710.

Carnap, Rudolf. *Philosophical Foundations of Physics*. New York and London, Basic Books, 1966.

Carnap, Rudolf. "Testability and Meaning". *Philosophy of Science*, volume 3, number 4, October 1936, pages 419-471 and volume 4, number 1, January 1937, pages 1-40.

Carroll, John W. "Nailed to Hume's Cross?". Chapter 2.1 of Theodore Sider, John Hawthorne and Dean W. Zimmerman (eds.), *Contemporary Debates in Metaphysics*. Oxford, Blackwell, 2008.

Carter, Angela. *The Passion of New Eve*. London, Virago, 1982.

Cassam, Quassim. *Self and World*. Oxford, Oxford University Press, 1997.

Cassam, Quassim (ed.). *Self-Knowledge*. Oxford, Oxford University Press, 1994.

Cei, Angelo, and Steven French. "Looking for structure in all the wrong places: Ramsey sentences, multiple realisability, and structure". *Studies in History and Philosophy of Science Part A*, volume 37, number 4, December 2006, pages 633-655.

Chalmers, David. "The Hard Problem of Consciousness". Chapter 17 of Max Velmans and Susan Schneider (eds.), *The Blackwell Companion to Consciousness*. Oxford, Blackwell, 2007.

Clark, Andy. "Reasons, Robots and the Extended Mind". *Mind & Language*, volume 16, number 2, March 2001, pages 121-145.

Clarke, Randolph. "Agent Causation and the Problem of Luck". *Pacific Philosophical Quarterly*, volume 86, issue 3, September 2005, pages 408-421.

Cohen, L. Jonathan. *An Essay on Belief and Acceptance*. Oxford, Clarendon Press, 1992.

Crick, Francis, and Christof Koch. "A Neurobiological Framework for Consciousness". Chapter 44 of Max Velmans and Susan Schneider (eds.), *The Blackwell Companion to Consciousness*. Oxford, Blackwell, 2007.

Cromwell, Oliver. *The Letters and Speeches of Oliver Cromwell with Elucidations by Thomas Carlyle*, edition in three volumes. London, Methuen, 1904.

Dahlbom, Bo (ed.). *Dennett and His Critics: Demystifying Mind*. Oxford, Blackwell, 1993.

Davidson, Donald. "Actions, Reasons, and Causes". Essay 1 of Davidson, *Essays on Actions and Events*, second edition. Oxford, Clarendon Press, 2001.

Davidson, Donald. "Agency". Essay 3 of Davidson, *Essays on Actions and Events*, second edition. Oxford, Clarendon Press, 2001.

Davidson, Donald. "Mental Events". Essay 11 of Davidson, *Essays on Actions and Events*, second edition. Oxford, Clarendon Press, 2001.

Dawkins, Richard. *Climbing Mount Improbable*. London, Viking, 1996.

Dayan, Peter, and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA and London, MIT Press, 2001.

Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Hassocks, Harvester Press, 1981.

Dennett, Daniel C. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA and London, MIT Press, 2005.

Dennett, Daniel C. *The Intentional Stance*. Cambridge, MA and London, MIT Press, 1987.

DeRose, Keith. "Solving the Skeptical Problem". *Philosophical Review*, volume 104, number 1, January 1995, pages 1-52.

Descartes, René. *Meditations*.

Dilthey, Wilhelm. "Das Verstehen anderer Personen und ihrer Lebensäußerungen". Part 3.1.2, pages 205-227, of Dilthey, *Der Aufbau der geschichtlichen Welt in den Geisteswissenschaften* (*Gesammelte Schriften*, volume 7). Stuttgart, B.G. Teubner Verlagsgesellschaft, 1958.

Dretske, Fred. "Contrastive Statements". *Philosophical Review*, volume 81, number 4, October 1972, pages 411-437.

Eilan, Naomi. "Perceptual Intentionality, Attention and Consciousness". Chapter 11 of Anthony O'Hear (ed.), *Current Issues in Philosophy of Mind: Royal Institute of Philosophy Supplement 43*. Cambridge, Cambridge University Press, 1998.

Elga, Adam. "Reflection and Disagreement". *Noûs*, volume 41, issue 3, September 2007, pages 478-502.

Enç, Berent. *How We Act: Causes, Reasons, and Intentions*. Oxford, Clarendon Press, 2003.

Evans, Gareth. *The Varieties of Reference*. Oxford, Clarendon Press, 1982.

Fine, Arthur. "The Natural Ontological Attitude". Chapter 1 of David Papineau (ed.), *The Philosophy of Science*. Oxford, Oxford University Press, 1996.

Fischer, John Martin. "Compatibilism". Chapter 2 of John Martin Fischer et al., *Four Views on Free Will*. Oxford, Blackwell, 2007.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, Cambridge University Press, 1998.

Frankfurt, Harry G. "Alternate Possibilities and Moral Responsibility". Chapter 8 of Gary Watson (ed.), *Free Will*, second edition. Oxford, Oxford University Press, 2003.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person". Chapter 16 of Gary Watson (ed.), *Free Will*, second edition. Oxford, Oxford University Press, 2003.

Frankish, Keith. *Mind and Supermind*. Cambridge, Cambridge University Press, 2004.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, Oxford University Press, 2007.

Fricker, Miranda. "The Value of Knowledge and the Test of Time". Chapter 7 of Anthony O'Hear (ed.), *Epistemology: Royal Institute of Philosophy Supplement 64*. Cambridge, Cambridge University Press, 2009.

Gert, Joshua. *Brute Rationality: Normativity and Human Action*. Cambridge, Cambridge University Press, 2004.

Gettier, Edmund L. "Is Justified True Belief Knowledge?". *Analysis*, volume 23, number 6, June 1963, pages 121-123.

Gigerenzer, Gerd. *Adaptive Thinking: Rationality in the Real World*. Oxford, Oxford University Press, 2000.

Ginet, Carl. "Reasons Explanation of Action: An Incompatibilist Account". Chapter 5 of Alfred R. Mele (ed.), *The Philosophy of Action*. Oxford, Oxford University Press, 1997.

Goethe, Johann Wolfgang. *Faust*.

Goldberg, Sanford C. *Anti-Individualism: Mind and Language, Knowledge and Justification*. Cambridge, Cambridge University Press, 2007.

Goldie, Peter. "Emotions, feelings and intentionality". *Phenomenology and the Cognitive Sciences*, volume 1, number 3, September 2002, pages 235-254.

Goldie, Peter. *The Emotions: A Philosophical Exploration*. Oxford, Clarendon Press, 2000.

Goldman, Alvin I. *Knowledge in a Social World*. Oxford, Clarendon Press, 1999.

Goldman, Alvin I. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford, Oxford University Press, 2006.

Goldstein, Sheldon. "Quantum Theory Without Observers – Part 1". *Physics Today*, volume 51, issue 3, March 1998, pages 42-46.

Grayling, A. C. *Truth, Meaning and Realism*. London, Continuum, 2007.

Haack, Susan. *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford, Blackwell, 1993.

Habermas, Jürgen. "The Language Game of Responsible Agency and the Problem of Free Will: How can epistemic dualism be reconciled with ontological monism?". *Philosophical Explorations*, volume 10, issue 1, March 2007, pages 13-50.

Hacking, Ian. *Representing and Intervening*. Cambridge, Cambridge University Press, 1983.

Haugeland, John. "Pattern and Being". Chapter 3 of Bo Dahlbom (ed.), *Dennett and His Critics: Demystifying Mind*. Oxford, Blackwell, 1993.

Heal, Jane. "Understanding Other Minds from the Inside". Chapter 3 of Heal, *Mind, Reason and Imagination*. Cambridge, Cambridge University Press, 2003.

Healey, Richard. *Gauging What's Real: The Conceptual Foundations of Contemporary Gauge Theories*. Oxford, Oxford University Press, 2007.

Hegel, G. W. F. *Wissenschaft der Logik*.

Heidegger, Martin. *Being and Time*.

Hendricks, Vincent F., and Duncan Pritchard (eds.). *New Waves in Epistemology*. Basingstoke, Palgrave Macmillan, 2008.

Hertz, Heinrich. *Untersuchungen über die Ausbreitung der elektrischen Kraft* (*Gesammelte Werke,* volume 2). Leipzig, Johann Ambrosius Barth, 1894.

Hitchcock, Christopher. "What Russell Got Right". Chapter 3 of Huw Price and Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford, Clarendon Press, 2007.

Hollis, Martin, and Steven Lukes (eds.). *Rationality and Relativism*. Oxford, Blackwell, 1982.

Hornsby, Jennifer. "Agency and Causal Explanation". Chapter 12 of Alfred R. Mele (ed.), *The Philosophy of Action*. Oxford, Oxford University Press, 1997.

Hume, David. *A Treatise of Human Nature*, second edition, edited by L. A. Selby-Bigge and revised by P. H. Nidditch. Oxford, Clarendon Press, 1978.

Hume, David. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, third edition, edited by L. A. Selby-Bigge and revised by P. H. Nidditch. Oxford, Clarendon Press, 1975.

Hurley, Susan L. *Consciousness in Action*. Cambridge, MA and London, Harvard University Press, 1998.

Hutto, Daniel D. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA and London, MIT Press, 2008.

Hutto, Daniel D. "Impossible problems and careful expositions: Reply to Myin and De Nul". Chapter 4 of Richard Menary (ed.), *Radical Enactivism: Intentionality, Phenomenology and Narrative: Focus on the philosophy of Daniel D. Hutto*. Amsterdam and Philadelphia, John Benjamins, 2006.

Hyman, John, and Helen Steward (eds.). *Agency and Action: Royal Institute of Philosophy Supplement 55*. Cambridge, Cambridge University Press, 2004.

James, Wendy. *The Ceremonial Animal: A New Portrait of Anthropology*. Oxford, Oxford University Press, 2003.

Johnson-Laird, Philip N. *How We Reason*. Oxford, Oxford University Press, 2006.

Kafalenos, Emma. *Narrative Causalities*. Columbus, Ohio State University Press, 2006.

Kant, Immanuel. *Critique of Pure Reason*. First edition (A), 1781, second edition (B), 1787.

Kant, Immanuel. *Gedanken von der wahren Schätzung der lebendigen Kräfte*. Ak.1: 1-181, 1747 (some sources give 1746, when the book was submitted to the censor).

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Ak.4: 385-463, 1785.

Kant, Immanuel. *Inaugural Dissertation (De Mundi Sensibilis atque Intelligibilis Forma et Principiis)*. Ak.2: 385-419, 1770.

Kant, Immanuel. *Von dem ersten Grunde des Unterschiedes der Gegenden im Raume*. Ak.2: 375-383, 1768.

Kaplan, Mark. "It's Not What You Know that Counts". *Journal of Philosophy*, volume 82, number 7, July 1985, pages 350-363.

Kierkegaard, Søren. *Kierkegaard's Journals and Notebooks*, volume 2, edited by Niels Jørgen Cappelørn et al., Søren Kierkegaard Forskningscenteret. Princeton, NJ, and Oxford, Princeton University Press, 2008.

Kierkegaard, Søren. *Søren Kierkegaards Skrifter*, volume 18, edited by Niels Jørgen Cappelørn et al., Søren Kierkegaard Forskningscenteret. København, Gads Forlag, 2001.

Kim, Jaegwon. "Mechanism, Purpose, and Explanatory Exclusion". Chapter 11 of Alfred R. Mele (ed.), *The Philosophy of Action*. Oxford, Oxford University Press, 1997.

Kim, Jaegwon. "Reasons and the First Person". Chapter 4, pages 67-87, of Jan Bransen and Stefaan E. Cuypers (eds.), *Human Action, Deliberation and Causation*. Dordrecht, Boston and London, Kluwer Academic Publishers, 1998.

King, Barbara J. *The Dynamic Dance: Nonvocal Communication in African Great Apes*. Cambridge, MA and London, Harvard University Press, 2004.

King, Ross D., et al. "The Automation of Science". *Science*, volume 324, number 5,923, 3 April 2009, pages 85-89.

Korsgaard, Christine M. *Self-Constitution: Agency, Identity, and Integrity*. Oxford, Oxford University Press, 2009.

Korsgaard, Christine M. *The Sources of Normativity*. Cambridge, Cambridge University Press, 1996.

Ladyman, James, and Don Ross, with David Spurrett and John Collier. *Every Thing Must Go: Metaphysics Naturalized*. Oxford, Oxford University Press, 2007.

Lamont, William (ed.). *Historical controversies and historians*. London, UCL Press, 1998.

Lange, Marc. *Natural Laws in Scientific Practice*. Oxford, Oxford University Press, 2000.

Leibniz, Gottfried Wilhelm. *Essais de Théodicée*.

Leibniz, Gottfried Wilhelm. *Monadology*.

Lemos, John. "Flanagan and Cartesian free will: a defense of agent causation". *Disputatio*, volume 2, number 21, November 2006, pages 69-90.

Lenk, Hans. "Humans as Meta-symbolic and Super-Interpreting Beings". Chapter 3 of Lenk, *Global TechnoScience and Responsibility*. Berlin and Münster, LIT Verlag, 2007.

Le Poidevin, Robin, and Murray MacBeath (eds.). *The Philosophy of Time*. Oxford, Oxford University Press, 1993.

Levin, Michael. "Gettier Cases without False Lemmas?". *Erkenntnis*, volume 64, number 3, May 2006, pages 381-392.

Lichtenberg, Georg Christoph. *Schriften und Briefe*, volume 2. München, Carl Hanser Verlag, 1971.

Lowe, E. J. *Personal Agency: The Metaphysics of Mind and Action*. Oxford, Oxford University Press, 2008.

McDowell, John. *Mind and World*, edition with a new introduction by the author. Cambridge, MA and London, Harvard University Press, 1996.

McFadden, Johnjoe. "Synchronous Firing and Its Influence on the Brain's Electromagnetic Field: Evidence for an Electromagnetic Field Theory of Consciousness". *Journal of Consciousness Studies*, volume 9, number 4, April 2002, pages 23-50.

MacIntyre, Alasdair. *After Virtue: A Study in Moral Theory*, third edition. London, Duckworth, 2007.

Mathiesen, Kay. "Collective Consciousness". Chapter 11 of David Woodruff Smith and Amie L. Thomasson (eds.), *Phenomenology and Philosophy of Mind*. Oxford, Clarendon Press, 2005.

Mele, Alfred R. (ed.). *The Philosophy of Action*. Oxford, Oxford University Press, 1997.

Menary, Richard (ed.). *Radical Enactivism: Intentionality, Phenomenology and Narrative: Focus on the philosophy of Daniel D. Hutto*. Amsterdam and Philadelphia, John Benjamins, 2006.

Menzies, Peter. "Causation in Context". Chapter 8 of Huw Price and Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford, Clarendon Press, 2007.

Montero, Barbara. "Varieties of Causal Closure". Chapter 8 of Sven Walter and Heinz-Dieter Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter, Imprint Academic, 2003.

Moore, Adrian W. *Points of View*. Oxford, Clarendon Press, 1997.

Mortimer, Ian. "The Death of Edward II in Berkeley Castle". *English Historical Review*, volume 120, number 489, December 2005, pages 1,175-1,214.

Musgrave, Alan. "NOA's Ark – Fine for Realism". Chapter 2 of David Papineau (ed.), *The Philosophy of Science*. Oxford, Oxford University Press, 1996.

Nagel, Thomas. *The Possibility of Altruism*. Oxford, Clarendon Press, 1970.

Nagel, Thomas. *The View from Nowhere*. Oxford, Oxford University Press, 1986.

Newman, Mark. "Ramsey Sentence Realism as an Answer to the Pessimistic Meta-Induction". *Philosophy of Science*, volume 72, issue 5, December 2005, pages 1,373-1,384.

Nietzsche, Friedrich. *Also Sprach Zarathustra*.

Nietzsche, Friedrich. *Die Fröhliche Wissenschaft*.

Nozick, Robert. *Philosophical Explanations*. Cambridge, MA, The Belknap Press, 1981.

Nussbaum, Martha. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*, revised edition. Cambridge, Cambridge University Press, 2001.

O'Brien, Lucy. *Self-Knowing Agents*. Oxford, Oxford University Press, 2007.

O'Connor, Timothy. "Agent Causation". Chapter 10 of O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. Oxford, Oxford University Press, 1995.

O'Connor, Timothy. *Persons and Causes: The Metaphysics of Free Will*. Oxford, Oxford University Press, 2000.

O'Hear, Anthony (ed.). *Current Issues in Philosophy of Mind: Royal Institute of Philosophy Supplement 43*. Cambridge, Cambridge University Press, 1998.

O'Hear, Anthony (ed.). *Epistemology: Royal Institute of Philosophy Supplement 64*. Cambridge, Cambridge University Press, 2009.

Olsson, Erik J. "The Place of Coherence in Epistemology". Chapter 8 of Vincent F. Hendricks and Duncan Pritchard (eds.), *New Waves in Epistemology*. Basingstoke, Palgrave Macmillan, 2008.

O'Shaughnessy, Brian. *The Will: A Dual Aspect Theory*, second edition, two volumes. Cambridge, Cambridge University Press, 2008.

Ozon, François (director). *Cinq fois Deux*. Film, 2004.

Papineau, David (ed.). *The Philosophy of Science*. Oxford, Oxford University Press, 1996.

Parfit, Derek. *Reasons and Persons*, corrected edition. Oxford, Clarendon Press, 1987.

Perry, John. "The Problem of the Essential Indexical". Chapter 10 of Quassim Cassam (ed.), *Self-Knowledge*. Oxford, Oxford University Press, 1994.

Persson, Ingmar. "Self-Doubt: Why We are not Identical to Things of Any Kind". Chapter 2 of Galen Strawson (ed.), *The Self?* Oxford, Blackwell, 2005.

Pettit, Philip, and Michael Smith. "Freedom in Belief and Desire". Chapter 20 of Gary Watson (ed.), *Free Will*, second edition. Oxford, Oxford University Press, 2003.

Pindar. *Odes*.

Pirsig, Robert M. *Lila: An Inquiry into Morals*, revised edition. Richmond, Alma Books, 2006.

Plato. *Phaedrus*.

Plato. *Theaetetus*.

Popper, Karl. *The Logic of Scientific Discovery*. London, Routledge, 2002.

Price, Anthony W. *Contextuality in Practical Reason*. Oxford, Clarendon Press, 2008.

Price, Huw, and Richard Corry (eds.). *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford, Clarendon Press, 2007.

Pritchard, Duncan. *Epistemic Luck*. Oxford, Clarendon Press, 2005.

Putnam, Hilary. *Reason, Truth and History*. Cambridge, Cambridge University Press, 1981.

Putnam, Hilary. "What is mathematical truth?". Chapter 4 of Putnam, *Mathematics, Matter and Method: Philosophical Papers, Volume 1*. Cambridge, Cambridge University Press, 1975.

Quine, Willard Van Orman. "Two Dogmas of Empiricism". Chapter 2 of Quine, *From a Logical Point of View*, second edition, revised. Cambridge, MA and London, Harvard University Press, 1980.

Quinton, Anthony. "Spaces and Times". Chapter 12 of Robin Le Poidevin and Murray MacBeath (eds.), *The Philosophy of Time*. Oxford, Oxford University Press, 1993.

Raffman, Diana. "Even Zombies Can Be Surprised: A Reply to Graham and Horgan". *Philosophical Studies*, volume 122, number 2, January 2005, pages 189-202.

Ramsey, William, Stephen P. Stich and Joseph Garon. "Connectionism, Eliminativism, and the Future of Folk Psychology". Chapter 9 of William Ramsey, Stephen P. Stich and David E. Rumelhart (eds.), *Philosophy and Connectionist Theory*. Hillsdale, NJ, Lawrence Erlbaum Associates, 1991.

Reid, Thomas. "Essays on the Active Powers of Man". In *The Works of Thomas Reid*, second edition. Edinburgh, Maclachlan, Stewart, & Co, 1849.

Roberts, Geoffrey. "Postmodernism versus the Standpoint of Action". *History and Theory*, volume 36, number 2, May 1997, pages 249-260.

Rovane, Carol. *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton, NJ, Princeton University Press, 1998.

Ruben, David-Hillel (ed.). *Explanation*. Oxford, Oxford University Press, 1993.

Rupert, Robert D. "Ceteris Paribus Laws, Component Forces, and the Nature of Special-Science Properties". *Noûs*, volume 42, issue 3, September 2008, pages 349-380.

Russell, Bertrand. "Why I Took to Philosophy". Chapter 4 of Russell, *The Basic Writings of Bertrand Russell*. London, George Allen & Unwin, 1961.

Ryle, Gilbert. *The Concept of Mind*. London, Hutchinson, 1949.

Schlosser, Markus. "Agent-causation and agential control". *Philosophical Explorations*, volume 11, issue 1, March 2008, pages 3-21.

Schlosser, Markus. *The Metaphysics of Agency*. PhD thesis, University of St. Andrews, 2006.

Schmidt, Michael, and Hod Lipson. "Distilling Free-Form Natural Laws from Experimental Data". *Science*, volume 324, number 5,923, 3 April 2009, pages 81-85.

Schopenhauer, Arthur. *The World as Will and Representation*.

Searle, John R. *Rationality in Action*. Cambridge, MA and London, MIT Press, 2001.

Seigel, Jerrold. *The Idea of the Self*. Cambridge, Cambridge University Press, 2005.

Shoemaker, Sydney. "Introspection and the Self". Chapter 7 of Quassim Cassam (ed.), *Self-Knowledge*. Oxford, Oxford University Press, 1994.

Sider, Theodore, John Hawthorne and Dean W. Zimmerman (eds.). *Contemporary Debates in Metaphysics*. Oxford, Blackwell, 2008.

Singer, Wolf. "Large-Scale Temporal Coordination of Cortical Activity as a Prerequisite for Conscious Experience". Chapter 47 of Max Velmans and Susan Schneider (eds.), *The Blackwell Companion to Consciousness*. Oxford, Blackwell, 2007.

Skinner, Quentin. *Visions of Politics: Volume 1, Regarding Method*. Cambridge, Cambridge University Press, 2002.

Smith, David Woodruff, and Amie L. Thomasson (eds.). *Phenomenology and Philosophy of Mind*. Oxford, Clarendon Press, 2005.

Smith, Michael. "The Structure of Orthonomy". Chapter 7 of John Hyman and Helen Steward (eds.), *Agency and Action: Royal Institute of Philosophy Supplement 55*. Cambridge, Cambridge University Press, 2004.

Sosa, Ernest. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford, Clarendon Press, 2007.

Sosa, Ernest, and Michael Tooley (eds.). *Causation*. Oxford, Oxford University Press, 1993.

Stevens, Wallace. "The Man with the Blue Guitar". *The Collected Poems of Wallace Stevens*. London, Faber and Faber, 1955, pages 165-184.

Steward, Helen. "Fresh Starts". *Proceedings of the Aristotelian Society*, volume 108, paper 11. Oxford, Blackwell, 2008.

Strawson, Galen. "Against Narrativity". Chapter 4 of Strawson (ed.), *The Self?* Oxford, Blackwell, 2005.

Strawson, Galen. *Mental Reality*. Cambridge, MA and London, MIT Press, 1994.

Strawson, Galen (ed.). *The Self?* Oxford, Blackwell, 2005.

Strawson, Peter F. "Freedom and Resentment". Chapter 1 of Strawson, *Freedom and Resentment and other essays*. London, Methuen, 1974.

Strawson, Peter F. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. London, Methuen, 1966.

Stueber, Karsten R. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. Cambridge, MA and London, MIT Press, 2006.

Sturgeon, Scott. "The Gettier Problem". *Analysis*, volume 53, number 3, July 1993, pages 156-164.

Taylor, Charles. *Sources of the Self*. Cambridge, Cambridge University Press, 1989.

Thrift, Nigel. "I Just Don't Know What Got into Me: Where is the Subject?". *Subjectivity*, volume 22, issue 1, May 2008, pages 82-89.

Tolstoy, Leo. *War and Peace*.

Turing, Alan M. "Computing Machinery and Intelligence". *Mind*, volume 59, number 236, October 1950, pages 433-460.

Tversky, Amos, and Daniel Kahneman. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment". *Psychological Review*, volume 90, number 4, October 1983, pages 293-315.

van Fraassen, Bas C. *The Scientific Image*. Oxford, Clarendon Press, 1980.

Velmans, Max, and Susan Schneider (eds.). *The Blackwell Companion to Consciousness*. Oxford, Blackwell, 2007.

von Wright, G. H. "On the Logic and Epistemology of the Causal Relation". Chapter 6 of Ernest Sosa and Michael Tooley (eds.), *Causation*. Oxford, Oxford University Press, 1993.

Walter, Sven, and Heinz-Dieter Heckmann (eds.). *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter, Imprint Academic, 2003.

Watson, Gary (ed.). *Free Will*, second edition. Oxford, Oxford University Press, 2003.

Weber, Max. *Wirtschaft und Gesellschaft (Grundriss der Sozialökonomik, III. Abteilung)*. Tübingen, J. C. B. Mohr (Paul Siebeck), 1922.

Wiggins, David. "Wittgenstein on Ethics and the Riddle of Life". *Philosophy*, volume 79, number 3, July 2004, pages 363-391.

Williams, Bernard. "Deciding to believe". Chapter 9 of Williams, *Problems of the Self*. Cambridge, Cambridge University Press, 1973.

Williams, Michael. *Problems of Knowledge: A Critical Introduction to Epistemology*. Oxford, Oxford University Press, 2001.

Williams, Michael. "Responsibility and Reliability". *Philosophical Papers*, volume 37, number 1, March 2008, pages 1-26.

Williamson, Timothy. *Knowledge and its Limits*. Oxford, Oxford University Press, 2000.

Wittgenstein, Ludwig. *On Certainty*, corrected edition, translated by Denis Paul and G. E. M. Anscombe. Oxford, Blackwell, 1974.

Wittgenstein, Ludwig. *Philosophical Investigations*, third edition, translated by G. E. M. Anscombe. Oxford, Blackwell, 2001.

Wittgenstein, Ludwig. *Tractatus logico-philosophicus*. Frankfurt am Main, Suhrkamp, 1963.

Woodward, James. "A Theory of Singular Causal Explanation". Chapter 10 of David-Hillel Ruben (ed.), *Explanation*. Oxford, Oxford University Press, 1993.

Woodward, James. "Causation with a Human Face". Chapter 4 of Huw Price and Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford, Clarendon Press, 2007.

Worrall, John. "Structural Realism: The Best of Both Worlds?". Chapter 7 of David Papineau (ed.), *The Philosophy of Science*. Oxford, Oxford University Press, 1996.