ABSTRACT

Rowland Stout

University College

D.Phil thesis

Trinity term 1991

The Teleological Explanation of Action

A different analytical approach to that of the standard causal theory to the explanation of action is proposed. It is argued that the most basic kind of explanation of action is teleological explanation in terms of external reasons. What this amounts to is that an action is the result of a causal process which adapts its results to whatever is objectively practically rational. Explanation in terms of psychological states depends on being able to make this externalist sort of explanation.

Central to this account is a theory of causal explanation which depends on the notion of a causal process. A causal process is a real entity distinct from an event. A phenomenon is causally explained when a description of the phenomenon is determined by a theoretical structure which represents how a process which results in the phenomenon works.

In teleological explanation, the theoretical structure is that of practical rationality. It is argued that this must be regarded as objective practical rationality. Only purposeful activity can be explained in this way. An account of evolutionary function is provided to show why it differs from this.

This account of teleological explanation, because it does not involve internal mental states, may be used to show how we attribute such states. An agent is essentially a teleological machine. Accounts of perception. beliefs and intentions are provided based on this.

Grammatical Note and Acknowledgement

I use the word 'they' rather than 'he' as my personal pronoun of indeterminate sex. I assume that this is new and not bad grammar.

I would like to thank my supervisor, David Charles, for the huge amount of help he has given me during my writing of this thesis.

Contents

Chapter 1

Teleological Explanation and the Causal Theory of Action

1.1 The Internalist Shift and The Causal Theory of Action

Explanation of action is characterized by the fact that the reason for an action justifies it at the same time as explaining it. Let us call such explanation *teleological*. (This is often misleadingly described as 'reason-giving' explanation, as if any explanation wasn't.) For example:

"Boris returned the jacket because he could not afford it."

"I watered the plant so that it would not die in the drought."

In these examples the justifying reasons are facts quite independent of the mental world of the subject. So we call this form of teleological explanation *externalist*. It is commonly supposed that the externalist form does not express the real, or at any rate the most immediate, explanation of action. According to this thought, the real reason that Boris returned the jacket was not that he could not afford it, but that he *thought* that he could not afford it. Perhaps the fact that he could not afford the jacket causally explains why he thought that he could not afford it. But, given that he *did* think that he could

not afford it, it is entirely irrelevant to the explanation of his action whether or not he really could afford it.

Let us describe this way of analysing externalist teleological explanations of action in terms of underlying internalist explanations as the internalist shift. According to the internalist shift, actions are always to be most immediately explained in terms of the thoughts, beliefs, desires, intentions, etc of the agent. Applying the internalist shift to the other example would lead us to suppose that the real explanation was as follows:

> "I watered the plant because I wanted it not to die in the drought and I believed that watering it was the way to achieve this aim."

The internalist shift does not *rule out* the correctness of externalist explanations of action. What it does require is that internalist explanations are (in some sense yet to be determined) more *immediate* and more *basic* than externalist explanations. But this still allows the possibility of externalist explanations which have non-immediate and non-basic roles. The best way to account for externalist explanations given the internalist shift is to claim that they are compound explanations; they have two parts - one of which is the immediate explanation of action in terms of beliefs and the other is the explanation of belief in terms of the facts.

The fact that Boris could not afford the jacket might explain why Boris *thought* that he could not afford it, and so *indirectly* explain why he returned it. According to this idea an externalist explanation

of Boris' behaviour can be attributed, but only in virtue of being able to attribute the more immediate internalist explanation.

This might all seem too obvious to be worth mentioning. Yet I argue that an important mistake is made if we accept this way of analysing the explanation of action. I think it has done for the philosophy of mind and action just what the equivalent internalist shift in epistemology has done for epistemology; i.e. lead to scepticism. So I deny the internalist shift, and show that by accounting for the teleological explanation of action in its externalist form it is possible to slice right through the philosophy of action and into the mind itself.

It is easy to find philosophers who fully endorse the internalist shift. It is clear that Hume was one. But Davidson is the best to quote since he is so explicit about it.

> *'Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of pro attitude towards actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind. ... Let us call this pair the* primary reason *why the agent performed the action. Now it is possible both to reformulate the claim that rationalizations are causal explanations and to give structure to the argument by stating two theses about primary reasons:*
> *1. In order to understand how a reason of any kind rationalizes an action it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason.*
> *2. The primary reason for an action is its cause.'*
> (Davidson 1980 pp 3,4)

Davidson's main concern is with the second of these theses - the causal thesis. But my objection is to the first alone - the conceptual thesis. In urging resistance to the internalist shift, I am not arguing that beliefs, desires, etc, *cannot* be reasons and causes for action. I think it is fairly clear that they can. What I deny is that an internalist explanation forms the core of every externalist explanation. I deny that we should account for externalist explanations *in terms of* internalist explanations, or attribute externalist explanations *in virtue of* the attribution of internalist explanations. What I am suggesting is that there is a type of teleological explanation that does not work with internal reasons, and that it is only *in virtue of* this *externalist* teleological explanation that internal states can explain behaviour.

So my argument concerns what should be the direction of the analysis concerning teleological explanation and mental states. But this only makes sense in the context of there being such a thing as analysis, and it having a direction; and such a context is by no means universally accepted. Many philosophers think that the notion of an analysis or an account is thoroughly bogus. Others may accept that there is a useful notion of analysis but deny that there is a useful notion of *direction* of analysis. They think of analysis as conceptual geography; i.e. the mapping out of symmetrical conceptual relations. According to this view, the role of analysis should be limited to providing elucidating connections between concepts.

So I must now come clean about the notion of analysis I will be working with throughout this thesis. This is particularly important since assumptions I make here with fairly minimal supporting argument determine the whole course of the thesis. For example, in chapter 6, I argue against the holism of the mental. This argument is only possible if one is working with an unholistic notion of conceptual analysis. So all results in the thesis should be regarded as conditional on the assumptions made about analysis.

Conceptual analysis is quite different from straightforward analysis of meaning. Its function is somehow to elucidate concepts and to investigate the structure of thought. So, conceptual analysis does not answer the question, 'What does x mean?' but rather it answers such questions as, 'In virtue of what is x true?' or, 'In virtue of what should we assert that x?' But it must do more than just provide correct answers to such questions. For, one true answer is, 'In virtue of the fact that x,' and another is, 'In virtue of the fact that y,' where y is a synonym of x. Neither of these answers takes us further than analysis of meaning.

I suggest that what we are really looking for is an informative description of the right way the concept should be applied. But sceptics of this approach to conceptual analysis question that such a description need be available. To begin with, they might argue that I am assuming that there is just one way of correctly applying a concept; yet clearly there is an indefinite number of ways.

The notion of conceptual analysis I want to work with is based on the assumption that there is something in common between all the ways of correctly applying a concept. We are finite creatures with limited time and limited capacities. We can completely learn how to use a concept and we can completely understand the use of a concept. So the correct way to use a concept must have some finite form. In much the same way there is an indefinite number of ways of correctly doing a throw-in in soccer. But, because we can learn in a finite time how to do a throw-in and we can adjudicate whether a throw-in is correct, there must be some common learnable and knowable core to these ways.

A threat to this conception might seem to come from Wittgenstein's arguments about rule following in the Philosophical Investigations (Wittgenstein 1958 secs 65-242). He argues that learning how to do something does not involve some sort of internalization of a self-contained formula of how to do it. For one thing there is an open-endedness about the rules of language. So, instead of being like soccer, it is like a game in which people start throwing a ball about and rules evolve, not through the development of formulae for the rules but just through the development of the practice of playing the game.

This analogy does no real damage to the notion of conceptual analysis; in fact it supports it. Analysing the rules of this open-ended game by describing how it should be played is a perfectly reasonable and perhaps important task. The open-endedness of the game means that there can be no complete and final specification of how the

game should be played. But an incomplete and provisional specification is still possible.

A similar response is possible to Wittgenstein's argument concerning how to add 2 to a number. His idea is that there can never be a complete explicit and self-sufficient specification of how this should be done. Every formula describing how to do something itself needs interpreting; i.e. it needs another formula describing how to interpret it.

But this is only an argument against the thought that behind every bit of language use there is a complete self-sufficient analysis that can explain why the use of that bit of language is correct or incorrect. It is not an argument against the possibility of saying something useful though incomplete in answer to the question, 'In virtue of what is the use of this concept correct?'

Another threat comes from the opposite direction. This is the argument that all this talk of ways in which a piece of language should be used, language games, etc, smuggles verificationism into the picture. According to this argument, analysis claims to dissect meaning, but does it by considering epistemological questions. By confusing epistemology with meaning-analysis, philosophers, so the argument goes, have been wrongly pushed into producing reductionist accounts.

But this does not bother me much either. I am explicitly not providing analysis of meaning. I am providing an account of how we should use bits of language. The only verificationist assumption here is that understanding language involves having a practical capacity.

Describing how such a capacity should work has no relation to reductionism.

Consider the causal theory of action in the light of these remarks. Let us start with the following formulation for the sake of argument.

> Activity resulting in some goal counts as intentional action directed to that goal if and only if the activity is caused in a non-deviant way by the intention to achieve that goal.

I think that this is quite true and obvious. Yet I deny the causal theory. This is because the statement above is not merely being asserted by the causal theorist; it is being presented as an account of what it is in virtue of that we correctly attribute intentional action. My criticism is that this statement does not show us (nor is it part of a larger account of) how we should attribute purposiveness to activity.

Let me illustrate the sort of analytical uselessness that I am trying to attribute to the causal theory. Suppose that the correct way of attributing deviance to a causal chain was by checking whether an intentional action was being caused through that chain by the appropriate intention. For, in that case, it would be impossible to tell whether activity was intentional by seeing among other things whether it was caused in a non-deviant way. The account would be circular, and therefore impossible to apply. This would make it useless as an account.

The causal theory needs an alternative way of accounting for deviant causal chains. And, as far as I know, it has not got one.

However, this is not my objection. I think that the causal account could make use of the notion of a causal process which I describe over the next few chapters to rid itself of the problem of deviant causal chains.

My objection is with the notion of *intention* in the causal theory. I do not think that the causal theory has access to an independent way of attributing intentions. In this thesis I argue that the way we should attribute intentions to a system depends on *already* being able to attribute intentional activity to it. If this is the case the causal account is useless.

It might seem that my claim is obviously false. Often we need to know whether a person intended to do something *before* we can tell whether what they did was intentional. But this misses the point. For, according to my account, what it is for a person to have some intention depends on how they *would* behave in an indefinite number of hypothetical situations. Usually we do not have to wait for a person's action before determining a person's intentions. So it is quite consistent to claim that *one* way of attributing intentional action depends on attributing the relevant intention, while claiming that in general intentional action can be attributed independently of any consideration of intentions, and that intentions are not rightly attributed independently of any consideration of intentional activity.

By putting in more structure to the causal theory I can show clearly where I disagree with it. First of all, according to the causal theory, purposiveness is to be understood in terms of the causal

teleological explanation of action. I.e. activity is purposeful when it is causally explained by justifying reasons. This is a common place for critics of the causal theory to dig in. But, I have no objection to this stage, whether the justifying reasons are supposed to be internalist or externalist. Opponents of the causal theory, like Melden and von Wright (Melden 1961, von Wright 1971), argue that logical connections between intentions and actions mean that intentions cannot be causes of actions. Davidson quite rightly denies this argument. But by concentrating on the causal thesis the real argument against the causal theory is overlooked.

Consider this quotation from Melden:

> *'It is impossible to grasp the concepts of motive and desire independently of an action. And, further, the sense in which a motive or a desire explains an item of conduct is altogether different from the sense in which, say, the presence of a spark explains the explosion of a mixture of petrol vapour and air.'*
> (Melden 1961 pp171,172)

I think that Davidson is right to criticize the thought expressed by the second sentence of this quotation. But, however right he may be in this respect, it just distracts attention from the important point to come out of Melden, which is expressed by the first sentence. That is that the logical connection between motive and action holds because one can only attribute motive in virtue of attributing action.

Another problem, as I see it, with the standard criticisms of the causal theory of action is that they seem to fully accept the internalist shift. Von Wright for example (von Wright 1971 pp 83,84)

argues that although there is a nomic relation between belief and action, the belief may be mistaken; so there is no nomic relation between fact and action. From this he concludes that externalist teleological explanation is not causal.

The mistaken assumption in this argument is that the causal relation betwen fact and action must depend on that between belief and action. (See the next two sections for more consideration of such arguments.)

Melden and von Wright end up by claiming that teleologically explaining something is just putting that thing in a certain context - the context of purpose. So they believe that teleological explanation of action is only to be understood in terms of purposiveness and not the other way around. Unfortunately, they conspicuously lack any other reasonable account of purposiveness.

Needing to assert that teleological explanations are not causal marks that a certain level of desperation has been reached. There is no objection to the notion of non-causal explanations. They may be what one is looking for when one says, 'I don't understand this, can you put it in context for me?' It is certainly true that one may want this when one does not understand someone's bodily movements or their actions. However, one may also say, 'I don't understand this, can you tell me why it has happened?' And in this case it is entirely inappropriate to provide a non-causal explanation. What one is asking to know here is what made that thing happen. One may ask *this* question of a piece of bodily activity or of an action, and then a

teleological explanation satisfies one. Therefore a teleological explanation is causal.

The next stage of the causal theory is that explanation of action is to be accounted for in terms of beliefs and desires (or intentions). This is the internalist shift which I am opposed to. Putting the account of purposiveness in terms of the causal explanation of action together with the internalist shift gives us the causal theory of action.

In addition to denying the internalist shift which is central to the causal theory, I argue that the causal theory has not got access to any independent way of attributing mental states like beliefs and intentions.

I do not account for explanation of action in terms of mental states. I have another way. This means that I can account for purposiveness without depending on the notions of mental states like intentions and beliefs. This leaves the way open for these mental states to be accounted for in terms of purposiveness and the teleological explanation of action. (This I do in chapter 6.) The following table shows in summary how my account differs from the causal theory of action. ($\phi \leftarrow \psi$ means ϕ is accounted for in terms of ψ.)

The Causal Theory

Purposiveness ← Teleological explanation of action ← Intentions/Beliefs

My Account

Intentions/Beliefs ← Purposiveness ← Teleological explanation of action

The position I advocate is not popular; there is almost no expression in the standard philosophical literature on action of a complete rejection of the internalist shift, despite the fact that the position would seem to be natural to many Wittgenteinian philosophers. What *can* be found is a complete rejection of the internalist shift in *epistemology*. It is now a respectable position that one's reasons for judgement might be the very facts manifesting themselves to one, rather than some intermediate internal representations of the facts in experience.

McDowell (see especially McDowell 1982) is an influential proponent of this line. He is also an influential proponent of a partial rejection of the internalist shift in the explanation of action. This position is associated with the so-called Kantian view in moral philosophy.

> *'To a virtuous person, certain actions are presented as practically necessary - as Kant might have put it - by his view of certain situations in which he finds himself. The question is whether his conception of the relevant facts weigh with him only conditionally upon his possession of a desire.'*
> (McDowell 1978 p14)

And the answer is no.

> *'It seems to be false that the motivating power of all reasons derives from their including desires.'*
> (McDowell 1978 p15)

But even McDowell seems to have made half of an internalist shift. Although he denies that desire must have a role in any complete teleological explanation of action, he explicitly accepts that belief must remain. For McDowell, it is the person's *view* of certain situations and his *conception* of the relevant facts rather than the situations or the facts themselves which justify the person's action.

My claim is more radical than this. For I claim that situations and facts justify a person's actions directly, and they do not weigh with him only conditionally upon his having a conception/view/belief of them. The usual debate in this area is between people like Davidson on the one hand who claim that beliefs and desires must be part of a complete teleological explanation and people like McDowell on the other hand who claim that only beliefs need be part of a complete teleological explanation. Although I will argue with McDowell against Davidson, I am not particularly interested in that debate. For I think that they are *both* mistaken in making any kind of internalist shift at all.

In the next two sections I consider why the internalist shift, even for a philosopher like McDowell, seems so compelling; and I try to undermine its attractiveness.

1.2 Arguments for the Internalist Shift

To begin with, consider the obvious argument *against* the internalist shift. The thought behind it is that internal reasons like beliefs, thoughts, desires, etc, cannot *justify* actions. At best they can only mitigate them. A justification of an action shows why that action is the right thing to do. This is an objective fact about the action that in no way depends on whether or not the agent believes it. The agent's beliefs may explain why an agent fails or succeeds in doing the right thing. But they do not *make* the action performed by the agent the right thing to do.

This should mean that, strictly speaking, internalist explanations are not teleological at all. It follows immediately that we should not account for the teleological aspect of explanation of action in terms of internalist explanations. Given that externalist explanations do justify action, then they must be the lynch pin in an account of teleological explanation.

One response to this might be the following. Perhaps external facts can justify actions in a stronger way than internal facts can (though see argument 3 later). But, *external facts cannot be the immediate reasons for action.* However good they are at justifying, they do not form an essential part of any explanation of action. So this ideally strong type of teleological explanation that involves immediate external reasons does not exist. On the other hand, internal reasons do justify action in some way, even if this way is just

mitigation. So, to this extent, internalist explanations are teleological. They might not be much, but they are all that we have.

If the argument is put this way, then the onus is seen to rest firmly on the proponents of the internalist shift to show that external facts cannot be immediate reasons for action. How might they do this? I consider five arguments for the internalist shift in this section, dedicating the next section to a consideration of a further argument which is the most interesting of all.

1. The first argument is that it is entirely mysterious how immediate causal explanation of action in terms of external facts could work. To this charge I present the next four chapters of this thesis. I hope to show how utterly natural and comprehensible such explanations are. However, I suspect that behind this charge of blank incomprehensibility lie one or more of the following arguments designed to show that immediate externalist teleological explanation is not just mysterious but straightforwardly impossible.

2. Consider Hume's argument that 'reason alone can never be a motive to any action of the will' (Hume's Treatise 2,3,3). This is designed to show that an agent's justifying reasons for an action must include desires (or other passions) as well as beliefs. If this argument works, then it is doubly bad for me. For I do not accept that immediate reasons for action must include beliefs let alone desires.

The basis of Hume's argument is that practical reasoning alone cannot take one from a state without a motivational element (a passion)

to a state with such a motivational element. Reason alone cannot inject passion; it can only redirect passion to its proper object.

The simplest way to fill in the argument for Hume's conclusion is to make the following observation: *unless an agent had some appropriate desire they would have no reason to act.* This observation is clearly true. It is supposed to follow equally clearly from this that the desire must be one of the agent's reasons to act. But it does not follow, as McDowell, taking his argument from Nagel (Nagel 1970 pp29,30), shows:

> *'the committment to ascribe such a desire is simply consequential on our taking him to act as he does for the reason we cite; the desire does not function as an extra component in a full specification of his reason.'*
> (McDowell 1978 p15)

The fact that there is a conceptual connection between having a desire and having a reason to act need not be explained by the fact that one has a reason to act in virtue of having the desire. It may be explained by the fact that one has the desire in virtue of having a reason to act; that is in virtue of the reason explaining or potentially explaining one's action.

So an externalist can accept that without a desire being present reason cannot motivate action. They do not need to assume that reason creates a motivational state from non-motivational states like plucking a rabbit out of a hat. If they assume that reason can yield action without there needing to be a motivational element in the reasons, then the motivational element can be attributed in virtue of that fact.

The onus is now on the externalist account to provide a successful account of how desires are to be attributed along these lines. I provide just such an account, for intentions at any rate, in chapter 6.

3. The Humean is going to feel cheated by this response. For, even if the externalist can successfully claim that we need not conceive of reason plucking *desire* out of the hat, they cannot deny that on this conception reason must pluck *action* out of the hat. How can a set of reasons not including a desire or other motivational state justify an action? Assuming that reason works in line with deductive logic on facts that are themselves motivationally neutral, it is impossible for reason to justify action. Some motivational input is required.

The brave externalist response to this argument is that objective facts need not be motivationally neutral. A fact is not independent of the concepts it is framed by. There seems to be no obvious objection to the possibility of deriving with reason justification of action from facts which already intrinsically contain action-justifying elements. Given a certain *way of looking at things* or a certain *conception of the facts*, reason can justify action without any purely subjective elements being included in any of the reasons.

This may be something like McDowell's position. But note there is one aspect of this position that is quite incidental to the task of fending off Humean internalism. That is the claim that morally- or motivationally-conditioned judgements about the world describe

objective facts. Whether or not this is true need not be decided for present purposes.

There are two ways in which an externalist can avoid grasping that nettle. First, they can simply deny that reason must work on objective facts, and require only that reason works on a *conception* of the facts. One can accept this requirement as a way of blocking the internalist argument, and remain neutral on the metaphysical question of whether beyond a conception of the facts there may lie a realm of unconceptualized facts. The problem with this strategy as I see it is that it still involves making the internalist shift for beliefs.

The second externalist strategy which bypasses the metaphysical question is to claim that *reason itself* may work in a way that is dependent on a way of looking at things. Rather than saying that the inputs to reason are concept-relative, we may say that reason itself is concept-relative. This is the line that I will take. In the next two chapters I argue that reasons in *any* kind of explanation do not justify according to deductive logic. Different explanations work according to different methods of rationalization. And it is the method of rationalization rather than the set of inputs to it that really represents the mechanism at work.

4. Now consider the arguments that suggest that immediate reasons for action must include *beliefs* rather than the facts that those beliefs are about. First, consider Williams' argument in his paper "Internal and External Reasons".

'Now no external reason statement could by itself *offer an explanation of anyone's action. Even if it were true (whatever that might turn out to mean) that there was a reason for Owen to join the army, that fact by itself would never explain anything that Owen did, not even his joining the army. For if it was true at all, it was true when Owen was not motivated to join the army. The whole point of external reason statements is that they can be true independently of the agent's motivations. But nothing can explain an agent's (intentional) actions except something that motivated him so to act. So something else is needed besides the truth of the external reason statement to explain action, some psychological link; and that psychological link would seem to be belief.'*
(Williams 1981 pp106,107)

If we take this argument at face value, it has the following form:

A. An external fact's truth is independent of whether or not an agent is motivated in line with it.

B. A reason must motivate the agent.

C. So an external fact cannot be a reason.

I assume, as I have done throughout, that by "reason" we mean "motivating reason". This means that the same external fact may in one situation be a reason and in another situation not be a reason. Whichever it is depends on whether the agent's activity is explained by it. So Williams' argument is invalid because there is the possibility that an external fact may or may not motivate an agent. When it does not motivate it is not a reason, but when it does motivate it is a reason.

Perhaps Williams is also arguing as follows:

D. A psychological link (belief) between fact and agent distinguishes cases where the agent is motivated in line with the fact from cases where the agent is not so motivated.

E. So the psychological link must be part of the explanation of the action.

McDowell (whom I am now arguing against) comes up with a similar argument.

> *'When we explain an action in terms of the agent's reasons we credit him with psychological states given which we can see how doing what he did, or attempted to do, would have appeared to him in some favourable light. A full specification of a reason must make clear how the reason was capable of motivating; it must contain enough to reveal the favourable light in which the agent saw his projected action.'*
> (McDowell 1978 pp14,15)

I assume that the second line of this quotation is supposed to follow from the first. But if this interpretation is correct, then a parallel argument to the one McDowell himself uses to deny that *desires* must figure in a complete specification of an agent's reasons should also work here. It is true that unless the agent had some appropriate psychological state (e.g. a belief) putting the projected action in a favourable light, they would have no reason to act. But it does not follow that the belief must be one of the agent's reasons.

For, (to rephrase an earlier McDowell quotation) the committment to ascribe such a *belief* is simply consequential on our taking the agent to act as they do, for the reason we cite; the *belief* does not function as an independent extra component in a full specification. According to this response, beliefs should be attributed in virtue of an agent's

activity or potential activity being explained (or potentially explained) by the relevant external facts.

This is an ambitious response which depends crucially on a certain sort of account of belief being possible. There are reasons why this response might seem less attractive in the case of beliefs than it does in the case of desires. The most obvious problem comes from the fact that we can have false beliefs. It is quite possible that there is no external fact that p to explain an agent's activity since p is false, yet we can still attribute the belief that p to the agent.

This would be an insurmountable problem to the externalist if they were claiming that we should attribute the belief that p to an agent solely in virtue of the fact that the fact that p teleologically explains the agent's behaviour. But the correct externalist account of belief is slightly more complicated than this, as is shown in chapter 6. In my account, the belief that p should be attributed in virtue of the fact that other external facts teleologically explain the agent's behaviour according to a method of rationalization that works as if p were true.

5. There is another mistaken reason for rejecting this response, which I suspect is implicit in both Williams and McDowell. This comes from a certain theory of explanation in which a complete specification of the reasons must form a logically sufficient condition of the thing being explained. According to this theory, if there is a valid explanation of some result in terms of a complete set of reasons, then it is

impossible that in some other situation all those reasons should be true and yet the corresponding result not happen.

It follows from such a theory of explanation that a full specification of a reason must incorporate the fact that the reason is motivating. But I propose a different theory of explanation in the next two chapters. In this theory, the full specification of the reason need not answer every question concerning the causing of the result. For, assuming a non-deductive model of explanation, the fact that the reason *is a reason* may be an extra piece of information. It is this extra fact rather than something incorporated in the reasons themselves that gives us the essential element of motivation. This is not to say that scientific enquiry can get no further towards specifying sufficient conditions for action. It is just to say that an explanation can be complete without being logically sufficient for what it explains.

1.3 The Argument from False Beliefs

I have left the biggest and best argument for the internalist shift till last. By giving it the grand title, The Argument from False Beliefs, I do not mean to suggest that it is the only argument exploiting the possibility of false beliefs - see e.g. argument 5 above. Nor do I mean that there is just one argument to be discussed under

this title. There are at least two quite separate species of argument to be discussed here; and each has an indefinite number of varieties.

What the title refers to is more a schema of an argument than an argument itself. Why I isolate this particular schema and give it this name is to bring out the illuminating parallel with another argument schema - The Argument from Illusion.

It is hard to find any explicit version of The Argument from False Beliefs in the literature, possibly because it seems too obvious to be thought of as needing such an airing. So let us feel towards it starting with a quotation from Davidson.

> *'Your stepping on my toes neither explains nor justifies my stepping on your toes unless I believe you stepped on my toes, but the belief alone, whether true or false, explains my action.'*
> (Davidson 1980 p8)

So it makes no difference to the correctness of an internalist explanation whether the belief is true or false. From this, it is supposed to follow that the externalist explanation, which only works if the belief is true, cannot be an immediate explanation of action.

When an agent acts on false beliefs, we cannot explain the action in terms of the facts but only in terms of those beliefs. But even when the beliefs are true that same internalist explanation works. The difference between cases where the beliefs are true and cases where they are not are all in the world outside the agent. In terms of the immediate explanation of the agent's behaviour there is no difference between the cases. So the same internalist explanation must lie at the

core of all explanations of action whether the beliefs are true or false.

The possibility of having a perfectly good explanation of action when the facts do not justify, forces us to explain action in terms of belief. Once we have this level of explanation in place the link from the facts to the beliefs is seen as not essential to the explanation of action. The second link is independent of the first (consider false beliefs). So, an account of explanation in terms of facts has two parts - an account of the formation of beliefs from facts and an account of the production of action from beliefs. Epistemology should deal with the first and Philosophy of Action should deal with the second.

I suggest that this is captured explicitly in the following argument:

A1 When an agent acts on the basis of false beliefs, the reasons for action are beliefs and not external facts.

A2 The explanation of the agent's action in terms of beliefs works just as well when the agent acts on the basis of true beliefs.

Therefore:

A4 There is no difference at the basic level in the agent's reasons for action between the case of acting on false beliefs and the case of acting on true beliefs.

Therefore:

A5 External facts do not figure in the most basic explanations of action, even in the case of acting on true beliefs.

A5 follows deductively from A1 and A4. But, as my numbering system suggests, there is something missing between A1, A2 and A4. Some hidden assumption needs to be brought out to make the argument valid. So there are three possible ways to try to block this argument. One might deny A1; one might deny A2; one might deny the step from A1 and A2 to A4 (i.e. deny whatever implicit assumption will fill the place of A3). I think that denying A1 is not the best move. Even if the expression of A1 might be improved upon, this would not affect the argument. A2 seems secure to me, though later I will briefly consider a threat to it. So the missing A3 is to be my target.

The gap at A3 allows that there might be different versions of this argument. I consider two. They can be simply distinguished by how they interpret the notion of the basic level in an agent's reasons. The first version works with a notion of explanatory basicness. The second works with a notion of epistemological basicness.

According to the first sense, I may explain the death of a tree in terms of the unusually mild winter, but the more basic reason is the honey fungus infestation which the mild winter allowed. And presumably there will be more basic explanations still (the strangulation of the xylem channels by the fungus, etc).

The more basic explanation (in this sense) is thereby analytically more basic also. For, the less explanatorily basic explanation holds only in virtue of the holding of the more explanatorily basic explanation. If internalist explanations of action are explanatorily more basic than externalist explanations of action, it follows

straightaway that we should account for the teleological explanation of action in terms of internalist explanations. Externalist explanations would have only a secondary role.

So one way of making the argument complete is by assuming that there can never be two different but equally basic explanations of the same thing. If we had to make a choice between an internalist and an externalist explanation in terms of explanatory basicness, we should have to choose the internalist explanation.

According to this version of the argument, the missing assumption is:

A3a If phenomenon P can be completely explained without reference to X, then X is not an essential part of a basic explanation of P.

This assumption captures the thought that there cannot be distinct equally basic explanations each completely accounting for the same phenomenon. The picture that leads to this assumption is of the explanation of action belonging to a series of Russian dolls with the explanandum in the centre. We can strip off the outer layers of explanation by changing the facts so that the beliefs are false. But, if the internalist layer remains, it will still explain the action. Because of the possibility of false beliefs in which the internalist explanation works and the externalist explanation fails, the internalist explanation must occupy a more central place than the externalist explanation.

But this picture of explanation is entirely bogus, and in as much as the Argument from False Beliefs depends on it, it is not correct. The mistake here is to think of all explanations as belonging to a one-dimensional hierarchy. In the next couple of chapters I provide an account of explanation which should make this clear. What will emerge out of this thesis is that internalist and externalist explanations belong to quite different categories of explanation and so cannot be placed in a single hierarchy. If this is the case A3a is straightforwardly false.

Principles like A3a are often blandly accepted as valid applications of Occam's razor. I.e. in this case: we should not over-populate the basic level of explanation with explanatorily redundant entities. But Occam's razor is a dangerous tool when applied blithely. I find it more plausible to follow the line developed in Strawson's *Individuals* (Strawson 1959), and say that the correct way to determine whether such and such a kind of entity exists is to see whether it is possible to identify and reidentify such things. This is in line with the weak verificationist assumption introduced earlier in 1.1 that understanding language is essentially a practical capacity. Knowing what it is for some sort of thing to exist is having a way of telling whether things of such a sort exist.

If this line is right, then there is no need to consider whether one *has to* postulate the existence of such things in order to explain the evidence (where the evidence is assumed to have some basic ontological standing). It is easy to show the danger of applying

Occam's razor. It is that one is likely to exclude entities which are in fact analytically more basic than the entities one retains. It is my hypothesis that this is just what has happened in the present case.

The second version of The Argument from False Beliefs interprets "basic" to mean "epistemologically basic". Something is epistemologically basic to a subject if they are directly aware of it: i.e. it is present to their consciousness or it is part of their mental contents, etc. Given this notion, we can understand The Argument from False Beliefs in the following way.

In the situation where an agent acts on false beliefs, beliefs and other mental states entirely constitute the agent's mental contents. No appropriate external facts belong to these mental contents. But, if the situation is changed in external ways only so that the beliefs become true, nothing is added to how it seems to the agent. So how it seems to the agent is still completely accounted for by the beliefs and other internal states constituting the agent's mental contents. The external facts still form no part of these contents. I.e. even in situations where the agent acts on true beliefs the external facts are not epistemologically basic to the agent.

To derive the internalist shift from A5 interpreted in this way we need an extra assumption.

A6 An agent's teleological reasons for action must be things that the agent is directly aware of (they must be part of the agent's mental contents).

I think that this assumption is fair enough. If the agent did not have some immediate mental access to the reasons for their behaviour, then the involvement of their agency in this behaviour would be brought into question. The immediate reasons for an agent's action must be *their* reasons.

So my denial of the internalist shift does not involve me in denying that an agent must have mental access to the immediate reasons for their actions. Instead, it involves me in denying that an agent cannot have mental access to external reasons for action. So, it is A4 that I deny.

The hidden assumption that makes the move to A4 valid according to this interpretation is the following:

A3b There cannot be distinct sets of mental contents each at the same time completely accounting for how it seems to the agent.

Like A3a this assumption might be justified in term's of Occam's razor, if Occam's razor were a valid tool. But I am assuming that we are not allowed Occam's razor. One other thing that might be behind this is the principle that qualitative indistinguishability determines the identity of mental phenomena. Call it the Qualitative Indistinguishability Criterion or QIC.

QIC If you are unable to distinguish A from something else (i.e. if they feel the same to you), then you are not directly aware of A (i.e. A forms no part of your mental contents).

QIC is a fair bit stronger than A3b. This is clear because if we assume QIC we do not even need to assume A2 in order to derive validly A4 and A5. All we need is the trivial assumption that an agent may be unable to distinguish a situation in which they are acting on true beliefs from one in which they are acting on false beliefs.

The argument rages over QIC. In favour, we have, for example, Blackburn:

> *'The* doppelganger *and empty possibilities are drawn up, as I have remarked, so that everything is the same from the subject's point of view. This is a legitimate thought-experiment. Hence there is a legitimate category of things that are the same in these cases; notably experience and awareness. Since this category is legitimate, it is also legitimate to ask whether thoughts all belong to it.'*
> (Blackburn 1984 p324)

Against this, we have, for example, McDowell:

> *'Notice how, instead of 'Everything seems the same to the subject', Blackburn uses locutions like 'Everything is the same from the subject's point of view'. This insinuates the idea - going far beyond the fact that there is a legitimate category of how things seem to the subject - of a realm of reality in which samenesses and differences are exhaustively determined by how things seem to the subject, and hence which is knowable through and through by exercising one's capacity to know how things seem to one, That idea seems fully Cartesian.'*
> (McDowell 1986 pp157,158)

This comment from McDowell seems to me to get it right. But it should be realized that for anyone who does not regard the label 'Cartesian' as worse than the plague, it does not constitute a knock-down argument against QIC. One further undermining thought about QIC

is that there is a principle that may well be genuine (it is a consequence of my analysis in chapter 6), which is very similar to QIC and a confusion between the two may be responsible for allegiance to QIC. This is what Evans describes as Russell's Principle.

> *'Russell held the view that in order to be thinking about an object or to make a judgement about an object, one must* know which *object is in question ... . Russell took this principle to require that someone who was in a position to think of an object must have a* discriminating conception *of that object - a conception that would enable the subject to distinguish that object from all other things.'*
> (Evans 1982 p65)

This principle can be phrased in a way to show its similarity with QIC.

> If you have no way of distinguishing A from something else, then you are not directly aware of A (i.e. A forms no part of your mental contents).

Russell himself thought that this sort of direct acquaintance could only be had with one's sense data. So he certainly conflated his Principle with QIC. But his Principle is not the same as QIC, because it is perfectly possible for two experiences to feel the same to a subject, even though the subject *does* have a way of distinguishing them. A subject may have a way of distinguishing between two things, but for some reason not apply that way, and so on that occasion be unable to distinguish between them. A subject may have a capacity without using that capacity.

For example, I may be unable to distinguish the experience of seeing Sarah from the experience of seeing her doppelganger Haras. Yet

I *have a way* of distinguishing between the two. I can check whether what I am seeing is Sarah rather than Haras: I can follow her back to her home here or on Twin Earth. So I do know how to tell whether it is Sarah I am seeing, and this is all that is required for me to have a discriminating conception of her. Russell's Principle (at least as described by Evans) does not force me to deny that I am seeing Sarah herself in this example. But, if we accept QIC, we must say that I am only directly aware of having an experience like that of seeing Sarah.

It is very instructive to see the parallels between the internalist shift in reasons for action and the internalist shift in reasons for belief. The classic argument for epistemological internalism is the Argument from Illusion. This exactly mirrors the Argument from False Beliefs.

B1 In illusion, mental contents do not include external objects.

B2 The mental contents of a subject in illusion are also present in veridical perception.

B3 There cannot be distinct sets of mental contents each at the same time completely accounting for how it seems to the subject. (Same as A3b.)

Therefore:

B4 There is no difference in mental contents between illusion and veridical perception.

(This intermediate conclusion can be reached more quickly with the aid of QIC and the trivial assumption that a subject may be unable to distinguish illusion from veridical perception.)

Combining B4 and B1 gives us:

B5 External objects never belong to mental contents.

One more Assumption is needed:

B6 A subject's justifying causes for perceptual judgement must belong to their mental contents.

Then we have derived the epistemological version of the internalist shift:

INT External objects cannot be justifying causes of perceptual judgements.

This interpretation of The Argument from Illusion owes much to McDowell (part III of McDowell 1982). His way of avoiding the conclusion of this argument is by invoking the so-called disjunctive account of appearances. (A version of the disjunctive account appears in Hinton 1973; then it is picked up by Snowdon 1981 before being found in McDowell 1982 and again in Snowdon 1990.)

> *'But suppose we say - not at all unnaturally - that an appearance that such-and-such is the case can be* either *a mere appearance* or *the fact that such-and-such is the case making itself perceptually manifest to someone. As before, the object of experience in the deceptive cases is a mere appearance. But we are not to accept that in the non-deceptive cases too the object of experience is a mere appearance, and hence something that falls short of the fact itself. On the contrary, we are to insist that the appearance that is presented to one in those cases is a matter of the fact itself being disclosed to the experiencer. So appearances are no longer conceived as in general intervening between the experiencing subject and the world.'*
> (McDowell 1982 p472)

The disjunctive account contains an explicit denial of the epistemological version of the internalist shift. The object of experience may be the "fact itself being disclosed to the experiencer". But in addition to this straight opposition to the internalist shift, the disjunctive account involves a way of undercutting The Argument from Illusion. For the disjunctive account shows us how we can avoid accepting B2. According to the disjunctive account, the mental contents of a subject in illusion are *not* present in veridical perception. Something may appear to be the case to a subject in virtue of one of two quite distinct objects of experience being present. In illusion one object of experience is present; in veridical perception another is.

This account involves a straight denial of QIC. I have no quarrel with this. But I think it may make sense to accept B2 even when one has no sympathy with QIC. The first thing to note is that McDowell implicitly accepts B3. This can be seen from his use of the word "the" in "the object of experience". Yet if he were to deny B3 he would have no need to deny B2 in order to block The Argument from Illusion. Instead of claiming that there is one object of experience in illusion and another object of experience in veridical perception, he could say that there was one object of experience in illusion and two objects of experience in veridical perception, one of which is shared with illusion and the other of which is the fact itself being disclosed to the experiencer.

These distinct objects of experience in veridical perception would belong to quite different categories, and so would not compete for

mental space any more than colleges compete for space with universities. Now, I do not wish to argue the point here for objects of experience in perception. I am not clear what this "mere appearance" is supposed to be, and it does not much matter for my purposes. For if we return to my original argument, The Argument from False Beliefs, it is more clear that the disjunctive approach is defective.

In this argument the relevant assumption is A2. This seems perfectly harmless and perfectly obvious. If we have an account that allows a subject to have direct experience of external facts, it seems quite unnecessary and implausible to insist that in such cases they do not have the same direct awareness of their beliefs as they do when their beliefs are false. There seems to be little intuitive problem with being directly aware that it is raining at the same time as being directly aware of believing that it is raining, without assuming that the latter has some more fundamental role in consciousness.

So the disjunctive account is not the best way to approach all the parallel arguments to The Argument from Illusion. It makes more sense to deny the parallel assumptions to A3 than to deny the parallel assumptions to A2. This is not to say that the disjunctive account of appearances is wrong. It is just that it is unnecessary to argue for it. (Note, for other purposes that the disjunctive account is put to it may not be unnecessary. For example, Snowdon uses it to deny that there is any conceptual connection between perception and the fact that some inner perceptual states are caused by outer objects (Snowdon 1981

and 1990). I suspect that something as strong as the disjunctive account is required to make this denial.)

One reason for looking for a way of denying The Argument from Illusion is to block the epistemological problems that seem to stem from it, in particular, scepticism about knowledge of the external world. There is nothing in the set of internal reasons for belief in the existence of the external world which justifies any belief stronger than that it seems as if there is an external world. So an internal system of justification cannot take us any further than the image of the world.

Exactly parallel considerations apply to The Argument from False Beliefs. I argue that the internalist shift leads to a problem in the philosophy of action very much like scepticism. The following is an argument that takes us from the internalist shift to the conclusion that actions themselves must be internal.

1. There is nothing in the set of internal reasons for action A that justifies anything more than an internal effort to achieve A.

2. So an internal justification cannot take us any further than the internal effort

3. Assume that every action is immediately justified only by a set of internal reasons. (This is the internalist shift.)

4. Then an action can be no more than an internal effort.

I find this conclusion disastrous. It seems to directly contradict the way we do talk about actions. Is it really a possibility that I perform exactly the action I perform in raising my

arm without moving my arm in any way? Hornsby is one philosopher who feels able to grasp the nettle of this consequence of the internalist shift. Her argument relies more directly on QIC than on the internalist shift, and goes as follows:

> *'If there is an action that is someone's intentionally moving his body, then that someone tries to move his body ... Moreover his trying to move his body is his moving it ... Trying to move the body if it is not an action is an internal event, possibly with external signs ... But it need make little difference to trying events* per se *whether or not they are actions ... So the tryings that are actions are also internal events.'*
> (Hornsby 1980 pp44,45)

Hornsby's claim that actions can be identified with tryings seems to me to be quite correct. But QIC infects the rest of the argument. According to QIC, the identificaion of tryings, because they are mental events, is determined by qualitative indistinguishability. One cannot necessarily tell the difference between a trying that does and a trying that does not involve body movements. So, according to QIC, they are the same event. This gives us Hornsby's line "But it need make little difference to trying events *per se* whether or not they are actions". It is this line that I deny, although I do accept that it is a direct result of the very assumption that may lead people to accept the internalist shift.

In both epistemology and the theory of action there is a major (perhaps insuperable) problem in getting an entirely internal system of reasons to justify anything outside the mental realm of the subject. Anyone who is suspicious of the assumption that epistemological

reasons must be internal, seeing this as a recipe for scepticism/idealism, should be equally suspicious of the exactly parallel assumption that reasons for action must be internal, seeing this as a recipe for Hornsby's idealistic theory of actions as internal events. Of course, this consideration is going to have no effect on anyone who feels that they can counter or come to terms with scepticism from the traditional epistemological position that the world of reasons is internal. But this position is becoming less and less attractive as Cartesianism is losing its stranglehold on philosophy.

1.4 Aristotle's Notion of Teleological Explanation

There are three claims that one can extract from Aristotle's writing about teleological explanation, each of which is, I argue, natural and correct, yet the combination of which yields an extremely powerful and unpopular idea. That idea is that value has a causal role in nature. The fact that Aristotle's notion of teleological explanation involves this idea suggests to most philosophers that there is something wrong with this notion. One or other of the central claims about teleological explanation needs to be disputed.

In this section I will introduce the three claims and indicate their plausibility. The idea that results from their combination will not be defended yet. However, I hope to have shown by the end of the

thesis how entirely natural and unmysterious it is that value has a causal role in nature.

The first claim concerns what essentially distinguishes teleological from non-teleological explanation. Although my main interest is with the teleological explanation of *behaviour*, a correct account of teleological explanation must show why it is natural to think of teleological explanations applying beyond behaviour to the products of evolution, the nature of artefacts, etc. There is something special about all such explanations. What is it? (Note that I am leaving open the possibility that though it is *rational* to think of teleological explanation applying to evolution it is nevertheless *wrong* to do so.)

One way of picking out teleological explanations is through the form of sentences describing them. The occurrence of phrases like 'in order to', 'for the sake of', 'so that', usually suggests that a teleological explanation is being given. However, this is a very unreliable test, and certainly does not amount to a definition. The explanation, 'I went to the bank because I had run out of money,' contains no superficial linguistic clue to its teleological nature. Indeed, the very same sentence might be providing a non-teleological explanation. For example, I ran out of money so I went to the cashpoint in order to get some more; but my card was consumed by the machine, so I stormed to the bank to complain to the manager. It is still true that I went to the bank because I had run out of money, but this is not the teleological explanation of my action. I was not

intending to get any money out of the bank, just to vent my frustration on the manager.

It seems that there must be something more substantial to teleological explanation than linguistic form. I think that this is captured in Aristotle's definition according to which a teleological explanation is explanation of something in terms of what that thing is for the sake of.

> *'For if a thing undergoes a continuous change toward some end, that last stage is actually that for the sake of which. [... But] not every stage that is last claims to be an end, but only that which is best.'*
> (Physics II 194a30-35)

If walking is good for health, and my walking can be explained by that reason, then it is teleologically explained. If a flower is for the sake of attracting insects and that is why a plant has one, then the presence of the flower is teleologically explained.

So the first claim that I want to extract from Aristotle is that something is teleologically explained only if it is explained in terms of the practical justification of that thing.

Claim 1: Teleological explanation is essentially explanation in terms of a justification.

I am deliberately going to leave the notion of practical justification open at this stage. There is an internalist version and an externalist version; and I do not want to prejudge this issue. What does seem to

be reasonably clear however is that *goals* must be involved in practical justification somehow.

At the very beginning I defined teleological explanation as a form of explanation in which the reason *justifies* the thing being explained. This was intended not as an exact stipulative definition: i.e. we shall call all and only such explanations teleological. Nor was it intended as a complete and accurate account of teleological explanation. Rather, it was just intended as a way of locating the concept: i.e. teleological explanation is *this* sort of explanation. The only claim I wish to be committed to at this stage is that the notion of justification or rationalization must be *part* of the correct account of teleological explanation.

It is easy to see why my definition as it stands is insufficient as a complete account of teleological explanation. For one thing, I would need to explain what the notion of explanation in terms of justification is. For another, it is quite conceivable that a reason for some phenomenon justifies the phenomenon, but only accidentally. The justification may be entirely unconnected with the causal explanation of the phenomenon. And, even if the justification is a causally essential part of the explanation, this may still be accidental if there is a deviant causal chain from justification to phenomenon.

For example, suppose that the dishes are dirty and need washing if they are not to be a health hazard. Also, to satisfy the internalists, suppose that I believe this and I want and intend to avoid a health hazard. Now, suppose that the fact of the health hazard posed by these

dishes (or my mental state concerning this fact) affects someone else in such a way that they threaten me with death if I do not do the dishes. Finally, suppose that I then do the dishes in order to avoid being killed (by the person not the germs).

In this example the fact of the health hazard (or my belief concerning that fact) justifies my doing the dishes and causally explains my doing the dishes. But this is not part of the *teleological* explanation of my activity. So the simple definition, if it were supposed to be a complete account of teleological explanation would fail.

This is one reason why existing externalist accounts of teleological explanation do not work. In these accounts, what makes an explanation teleological is that the causal reason has some future-directed aspect to it. Taylor's account is an example of this.

> *'To offer a teleological explanation of some event or class of events, e.g. the behaviour of some being, is, then, to account for it by laws in terms of which an event's occurring is held to be dependent on that event's being required for some event.'*
> (Taylor 1964 p9)

Larry Wright adapts this account.

> *'S does B for the sake of G iff*
> *(i) B tends to bring about G*
> *(ii) B occurs because (i.e. is brought about by the fact that) it tends to bring about G.'*
> (Wright 1976 p39)

Both these accounts characterize teleological explanation as explanation with a special sort of causal reason. But if there is a deviant causal chain from reason to result, that reason is no longer the *teleological* reason for the result. So these accounts fail to characterize teleological explanation. A correct account must include some special notion of *how* a teleolgical reason causes its result.

There is a worse problem yet with such accounts. The notion of justification implicit in them is too weak to capture the desired notion of teleological explanation. That an event is required for some end or tends to bring about some end does not justify that event unless the end is justified itself. The fact that in a teleological explanation the end must be a real *goal* is left out of these externalist accounts. By not requiring that there be a normative element to the characterization of teleological explanation, these accounts fail to capture the idea of a teleological explanation being an explanation in terms of *justification*.

The example of water in a U-shaped tube should make this clear. (See Woodfield 1976 p83.) When the water levels in the two arms of the tube have been unbalanced, the level in one arm of the tube drops because its dropping tends to equalize the two water levels. According to Wright's account, this is a teleological explanation. But if this is allowed to count as a teleological explanation, it looks as if teleological explanation is not the interesting and important notion that Claim 1 suggests it should be. Indeed, such an account, by failing to capture the point of Claim 1 fails to be an account of teleological

explanation at all. (An explanation of why Claim 1 is not satisfied by this account is given in 4.3.)

So I want to distance my account form the standard externalist accounts. I think that the normative aspect that they miss out is essential to a useful notion of teleological explanation. On the other hand, in as much as they are externalist, I think they are right and also capture another element of Aristotle's notion of teleological explanation.

Aristotle equates practical justification with means-ends justification. This suggests that he has in mind the externalist form of teleological explanation as central. It is certainly safe to attribute to him the view that teleological explanation works in cases where no internalist shift is even possible - i.e. in nature rather than in thought.

> *'Again, some of the former class [things that come to be for the sake of something] are in accordance with intention, others not. ... (Things that are for the sake of something includes whatever may be done as a result of thought or of nature.)'*
> (Physics 196b15-25)

This means that at the very least Aristotle has a notion of teleological explanation which is externalist and cannot be reduced to an internalist form. If general internalism is defined as the position according to which all teleological explanations are basically internalist, then Aristotle was not an ascriber to general internalism. This is the second Aristotelian claim that I wish to endorse.

Claim 2: There is a central notion of teleological explanation that is externalist.

I would also endorse a stronger claim than this; i.e. that *the* central notion of teleological explanation is externalist. However, one has to be very careful in drawing too strong an interpretation of Aristotle in this matter. There are suggestions in Aristotle's work on acrasia among other things that he is assuming an internalist shift in the explanation of action (see e.g. Charles 1984). If this is right then Aristotle must be interpreted as holding a position intermediate between general internalism and general externalism.

One way he might do this is to accept a mixed strategy, according to which there are different kinds of teleological explanation, those with the externalist form basic and those with the internalist form basic. This is the position advocated by Woodfield. Woodfield distinguishes functional teleological descriptions from purposive teleological descriptions. The former describe things happening because they *are* good. The latter describe things happening because they *are believed to be* good. According to this position it must be the linguistic form of teleological explanation that ties together the different sorts rather than anything more real.

> *'although functional and purposive explanations are assessed by completely different criteria, the similarity in their underlying grammatical structures explains why they are bracketed together under the same label.'*
> (Woodfield 1976 p206)

Whether or not there is any evidence elsewhere for Aristotle's having accepted the internalist shift, I do not think that his account of teleology in The Physics presupposes such acceptance. Woodfield interprets Aristotle as having two quite separate notions of teleological explanation (corresponding to Woodfield's purposive and functional teleological descriptions) which Aristotle then outrageously confuses. He quotes this piece of Aristotle.

> *' "That for the sake of which" tends to be what is best and the end of things that lead up to it. (Whether we call it "good" or "apparently good" makes no difference.)'*
> (Physics 195a25-27)

Woodfield assumes that the distinction between good and apparently good must mark a distinction between externalist and internalist forms of teleological explanation. But in fact it is quite possible to read this quotation as entirely consistent with *general* externalism. A general externalist has to accept that teleological explanations can be had even when the goodness of the goal is only apparent. But as I show in chapter 5, this does not mean that they must identify "apparently good" with "believed to be good". Explanation in terms of apparently good can be understood in terms of explanation in terms of the good with a certain kind of error allowed for. Woodfield's interpretation of Aristotle assumes that the possibility of teleological explanation working in the context of a mistake about what is good makes it unavoidable to accept the internalist shift. If this were so, then it follows from the fact that Aristotle recognizes such a

possibility that he must be making the internalist shift. But if I can show in chapters 5 and 6 that general externalism is consistent with this possibility, then we cannot jump to the conclusion that Aristotle was not generally externalist.

An alternative to all these positions is simply to refuse to accept that either the externalist or the internalist form is more basic. That Aristotle might support such a position is suggested by the unimportance which he attributes to the distinction. However, whether or not this is Aristotle's position, it is not one that *we* can now hold having seen the importance of the distinction. (For the claim that there is an important shortcoming in Aristotle's account of teleological explanation here see Charles forthcoming.)

The third claim about teleological explanation that I derive from Aristotle is the following:

Claim 3: Teleological explanation is a kind of causal explanation.

Aristotle seems very explicit about this.

> *'"Why is he walking about?" We say: "To be healthy", and having said that, we have assigned the cause. The same is true of the intermediate steps which we brought about through the action of something else as means towards the end.'*
> (Physics 194b30-36)

This is open to the obvious counter-interpretation that Aristotle means by 'cause' just 'explanation'. A 'formal cause' is a special kind of explanation rather than a special kind of causation. So why

shouldn't a teleological cause be also? (See e.g. Annas 1982.) However, it is becoming more usual nowadays to interpret Aristotle as regarding teleological explanation as a species of causal explanation after all. (See e.g. Cooper 1987 and Gotthelf 1987.) As long as one does not regard causal teleological explanation as such an absurd notion that Aristotle could not possibly have had it, it seems wise to take him at face value when he says that intermediate steps are *brought about* through the action of something else as means to ends.

The claim itself causes few problems for most people. The thing that does worry people about Aristotle's attitude to Claim 3 is his extra assumption that teleological causal explanation is *distinct* from efficient causal explanation. There is a feeling that the causal relations in the universe must be exhausted by the efficient causal relations, and that to suggest anything else is to invoke magic.

But this feeling may just be another case of Occam's razor getting into situations where it does not belong. One does not have to assume that a teleological explanation is independent of efficient causal explanation in accepting this Aristotelian assumption. A teleological explanation may occupy the same causal space as efficient explanations without being eliminable in favour of the efficient explanations. 'But if something can be explained without invoking teleology, isn't the teleological explanation redundant?' queries the Occam's razorist. And the response, as before, is to insist that this has nothing to do with the question of whether there are teleological causal explanations distinct from efficient causal explanations.

It is important to realize that this point about teleological explanation being distinct from efficient causal explanation is not really an *extra* assumption at all. In fact it follows from Claims 1, 2 and 3. If we treat Claim 1 seriously, it follows that teleological explanation is essentially normative. Given Claim 2, there is the possibility of irreducibly externalist teleological explanation. In such a case the element of justification cannot be accounted for in terms of efficient internalist causal explanation. But there seems no other possible way that this normative element can be fleshed out in non-normative terms.

Few philosophers since Aristotle have been willing to accept all three claims about teleological explanation. Externalists generally cannot stomach Claim 1 in its full strength. Internalists reject Claim 2, and for those areas where they cannot deny that the teleological explanation is externalist, they either reject Claim 1 themselves (saying that such explanations are not *really* teleological) or they reject Claim 3 (this is Woodfield's strategy).

It is easy to see why this is. Putting the three claims together yields the conclusion that things can be causally explained in terms of why they are justified, where this notion of justification cannot be fleshed out in terms of internal mental states in all cases. To say that something is justified is to make a value judgement concerning it. So the Aristotelian notion of teleological explanation involves values having a basic causal role in nature.

This is also the notion of teleological explanation I am working with. One of the tasks of this thesis is to show how natural and unmysterious this notion really is. I should note however that I am not committed to accepting Aristotle's idea of where such externalist teleological explanation applies. As I argue in chapter 4, I think it only applies to the products of evolution in a rather strained sense. For me, the central role for this sort of explanation is with action, although this was just the place where there seems to be some doubt about whether Aristotle would apply it.

1.5 Teleological Theories

One further aspect of Aristotelian teleological explanation is important. This is Aristotle's claim that the essential nature of a wide variety of things is determined by how they are teleologically explained - i.e. by what they are for the sake of. Now my principal interest here is with action; and concerning action this claim can be accepted by internalists and externalists alike. It is essential to the nature of action that it is purposive, and teleological explanation is essential to the nature of purposiveness.

In line with my notion of what an *account* of (e.g.) purposive action is (see 1.1), I would want to interpret this general claim about the role of teleological explanation in essential natures in a special way. The following is no more than a programmatic sketch of how this

might be done. I regard the important analytical question here as, 'In virtue of what should we say that such and such a sort of thing exists?' My proposed answer is, 'In virtue of identification of some aspect of its essential nature.' (This is in line with e.g. Wiggins' theory of individuation - Wiggins 1980.) So I recommend using essences in an account of how certain sorts of things are identified.

Applying Aristotle's claim about teleological explanation constituting things' essential natures to the analytical question, 'How should we identify such and such a kind of thing?' gives the following accounts. The way to identify the existence of an action is to tell that some bodily activity is teleologically explained in a certain way. Similarly the way to identify the existence of an artefact (e.g. a hammer) is to tell that the presence of a piece of metal attached to a piece of wood is teleologically explained in a certain way. I.e. it is there in order to bang home nails. Similarly the way to identify the existence of a body organ (e.g. a lung) is to tell that the presence of a certain piece of body tissue is teleologically explained in a certain way. I.e. it is there in order to aerate the blood.

Out of the underlying entities emerge these higher order entities when the underlying entities are teleologically explained.

Such teleological accounts can be seen as the culmination of a series of philosophical attempts at accounting for the identification of such entities. First came the intrinsicalist accounts. According to intrinsicalism, given bodily activity with certain intrinsic characteristics, we can identify an action. Given a piece of matter

with certain intrinsic characteristics (shape, hardness, etc), we can identify a hammer. Given a piece of body tissue with the right shape, size, structure, etc, we can identify a lung.

Then came causalist accounts, according to which, given bodily activity with certain causal properties (either input or output), we can identify an action. Given a piece of material which can cause nails to penetrate wood when someone swings it at them, we can identify a hammer. Given a piece of body tissue which can cause the aeration of the blood under certain causal conditions, we can identify a lung.

And finally come the teleological accounts, according to which, given bodily activity which can be teleologically explained in a certain way, we can identify an action. Given a piece of material whose presence can be explained in terms of the fact that it is good for banging in nails, we can identify a hammer. Given a piece of body tissue whose presence in the species can be explained in terms of its aerating the blood, we can identify a lung.

A very attractive suggestion is that this same development works for the account of mental states. Let us consider this suggestion first with respect to a materialist view of mental states.

An intrinsicalist materialist account identifies mental states in terms of intrinsically (i.e. physically in this case) characterized brain states. For example, 'Pain just is stimulation of C-fibres'. The failure of this account was initially put down to the failure to recognize the importance of the *causal role* of mental states. If a stimulation of C-fibres was not characteristically the result of a

damaging causal input and did not itself usually result in behaviour which stopped this input, it could not be identified with pain.

So, functionalism, a version of causal materialism, emerged. Here a certain mental state is identified with whatever brain state has a certain causal role (with inputs, outputs and other mental states) which corresponds with that mental state in a good psychological theory. If there is a brain state which has the causal role of belief that the sun is shining, then that belief can be identified as present.

The failure of functionalism may be put down to its failure to recognize the constitutive role of the ideal of rationality.

> *'Any effort at increasing the accuracy and power of a theory of behaviour forces us to bring more and more of the whole system of the agent's beliefs and motives directly into account. But in inferring this system from the evidence, we necessarily impose conditions of coherence, rationality and consistency. These conditions have no echo in physical theory, which is why we can look for no more than rough correlations between psychological and physical phenomena.'*
> (Davidson 1980 p231)

The logical relations between contents which must be incorporated in any theory on which attribution of mental states is based cannot be incorporated in the sort of physical theory that functionalism relies on. A finite number of ad hoc constraints can be incorporated: but rationality as such cannot be so easily captured. (This is argued for in McDowell 1985.)

The belief about a tree and the belief about the collection of molecules making up the tree do have different causal roles, but these

differences will be represented only in a theory that incorporates logic somehow.

The only kind of causally explanatory theory that gives a constitutive role to rationality is a teleological theory. According to Claim 1 in the previous section, a teleological explanation explains something in terms of that thing being rationally justified. So a theory that identifies a mental state in terms of its *teleological* role will be able to distinguish between a belief about a tree and a belief about the collection of molecules that make up the tree.

Teleological materialism is the view that, given brain states which are teleologically explained in a certain way, mental states can be identified. If the function of having a certain brain state is that the organism can thereby distinguish what is in front of it as a tree rather than some other kind of object (the ability to make this distinction being good for its survival prospects), then we can attribute the mental state of seeing a tree to this organism. This is the sort of position adopted by Millikan (1984), Papineau (1987) and a growing band of other disenchanted functionalist philosophers.

The materialist basis of this series of accounts might be criticized by a subjectivist on the grounds that *what it is like* to have one of these mental states is not captured by any materialist account. An *intrinsicalist* subjectivist account would identify mental states in terms of intrinsically (qualitatively in this case) characterized states of consciousness. This does not seem like much of an account at all since the problem of identifying qualitatively

characterized states of consciousness is just the same as the problem of identifying mental states. However, the account could still be seen to be useful as an account of *content-bearing* mental states. According to this account, we can identify a mental state with a certain content when we can identify a state of consciousness having a certain qualitative nature (of resemblance perhaps). This is the position of the British Empiricists.

The causalist and teleologist versions of subjectivism are much like their materialist counterparts. A causalist subjectivist argues that it is not enough for an experience to have the qualitative feel of being of a table, it must also be caused by a table if the subject is to be attributed with a perception of the table. A teleologist subjectivist argues that for the experience to have *that* content it must in addition be explained teleologically. In particular, the experience should have the function of enabling the subject to discriminate the present situation from situations in which there is no table.

Another criticism of the materialist basis of causalist and teleological accounts of mental states comes from Dennett's intrumentalism (e.g. Dennett 1987). Dennett points out that it does not make any difference to the identification of a mental state what physical brain state is playing the appropriate causal (or teleological) role. But then it does not make any difference either if there is no physical state at all playing the role as long as there seems to be one.

I will not pursue either the subjectivist or the instrumentalist alternatives to teleological materialism. For, all three accounts suffer from the following criticism. Assume that the Davidsonian criticism of the causalist accounts works. So, rationality imposes a general constraint on any theory of behaviour by which we can attribute mental states. This constraint cannot be captured in a non-teleological theory. The move now made by Millikan et al is to say that an evolutionary explanation of the physical state allows us to involve rationality at the centre of the account. (This move has obvious counterparts for subjectivism and instrumentalism.) The problem is that we are not looking for evolutionary rationality; we are looking for human rationality.

Evolutionary rationality works according to means-ends analysis on the top-level goal of propogation. It yields descriptions of what behaviour mechanisms an individual should have in order to survive and propogate. As such it simply does not have the resources to interpret human behaviour directed to human goals. Millikan's theory depends on the extraordinary claim that a person's rationality is an aspect of evolutionary rationality.

It should be admitted that I have not yet provided a knock-down argument against using the resources of evolutionary explanation in some way to *help* fix the content of mental states. But perhaps such an argument is available. Consider Millikan herself raising the problem. Suppose a cosmic accident produces a physical replica of you.

> *'that being would have no ideas, no beliefs, no intentions, no aspirations, no fears, and no hopes. (His non-intentional states, like being in pain or itching, may of course be another matter.) This is because the evolutionary* history *of this creature would be wrong. For only in virtue of one's evolutionary history do one's intentional mental states have proper functions, hence does one mean or intend at all, let alone mean anything determinate. ... That being would also have* no liver, no heart, no eyes, no brain, *etc. For the categories "heart", "liver", "eye", "brain", and also "idea", "belief" and "intention" are proper function categories, defined in the end by reference to long-term and short-term evolutionary history, not present constitution or disposition. Were this not so, there could not be* malformed *hearts or* non-functioning *hearts nor could there be* confused *ideas or* empty *ideas or* false *beliefs, etc. Ideas, beliefs and intentions are not such because of what they do or could do. They are such because of what they are, given the context of their history,* supposed *to do and of how they are supposed to do it.'*
> (Millikan 1984 p93)

Millikan's point about hearts, livers, etc may be right, but her extension of this thought to mental states is at best speculative. She has a theory to account for evolutionary mistakes. But it does not follow that the possibility of mistaken mental states must be accounted for evolutionarily also. Indeed, I think it is quite easy to show that evolutionary explanation has nothing to do with mental content.

The following argument is based on what Evans describes as Russell's Principle (Evans 1982 ch4).

RP: If one has no way of distinguishing A from A' then A forms no part of one's mental content.

RP is consistent with A being part of one's mental content and one *failing* to distinguish A from A'. But what is ruled out is that A forms part of one's mental content and one has no access to a way of

getting the distinction right. So, attribution of content is constrained by what discriminatory capacity the subject has. Combine this with the following two fairly trivial assumptions about teleological materialism.

1. According to teleological materialism, the fact that the content of some mental state is A rather than A' may be entirely due to the fact that A rather than A' figures in the evolutionary teleological explanation of one's states.

2. One can correctly be described as having A rather than A' as the content of this mental state even when one has no conscious access to the evolutionary basis for the distinction (i.e. even though one has no way of distinguishing A from A' in the evolutionary explanation of one's state).

It follows from these three assumptions that teleological materialism is wrong. A teleological materialist may return with the argument that the specification of a subject's discriminatory capacity depends on the teleological explanation of that capacity. But even if this is correct, it is too weak to block the argument. So, one may grant to the teleological materialist that whether a capacity should be described as the capacity to distinguish A from B rather than the capacity to distinguish A' from B' is to be determined by the teleological function of that capacity. But the argument does not concern this capacity, but the capacity to distinguish A from A'. And by hypothesis, this capacity is not present at all whatever the description.

Let us assume some version of Russell's Principle. I.e. the content of a subject's mental state depends on the subject having a

discriminatory capacity with respect to that content. We can also allow that the correct identification of the discriminatory capacity must depend on teleological explanation somehow. According to teleological materialism, the correct specification of the discriminatory capacity depends on the evolutionary explanation of that capacity. But there is a much simpler and better motivated way of connecting the identification of a discriminatory capacity and teleological explanation. A discriminatory capacity results in behaviour whose rationality depends on the distinction being correct. So I have the capacity to distinguish the object in front of me as a tree rather than anything else if I have the capacity when the rationality of my behaviour depends on whether or not the object is a tree to behave accordingly. When a discriminatory capacity results in such discriminatory bahaviour, that behaviour is explained in terms of the fact that it itself is justified somehow. The behaviour is teleologically explained.

If this line of thinking is right, then it is wrong to account for content-bearing mental states in terms of the teleological explanation of physical capacities. Instead, mental events should be accounted for in terms of the teleological explanation of the *behaviour* which results from the physical capacities. So, what I am endorsing can be described as teleological *behaviourism*.

According to intrinsicalist behaviourism, a person should be described as angry in virtue of their behaviour being angry. An improvement on this was causalist behaviourism, according to which, a

person should be described as wanting to stay dry in virtue of their behaviour being caused by a mechanism (disposition) that results in staying-dry behaviour. On this view, since the mental state is not constituted by the behaviour, it makes sense to attribute it even when the staying-dry behaviour does not come about, so long as the disposition to stay dry can be identified.

The standard objection to causalist behaviourism is the same as the standard objection to causalist materialism. The constraints of rationality have a role in the correct identification of mental states which is not capturable in a causalist theory. Teleological behaviourism captures this role centrally. Moreover, since teleological behaviourism, unlike teleological materialism, is not restricted to a natural selection account of teleological explanation, it captures the very constraints of rationality that we operate with. There will be no need to attempt the heroic task of accounting for human rationality in terms of evolutionary rationality.

In my argument against teleological materialism I have assumed that the teleological materialist must regard teleological explanation as evolutionary explanation. But why shouldn't they work instead with an internalist type of teleological explanation? According to this possibility, a subject has a certain mental state if that state figures in a holistically ascribed set of mental states which together would internalistically teleologically explain the person's behaviour. This position seems more like the natural successor to functionalism than Millikan's teleological materialism does. It incorporates Davidson's

point about the constitutive role of rationality more centrally. It is also very close to my teleological behaviourism. Really the only difference is that in this version of teleological materialism, teleological explanation is regarded as basically internalist, and in my version of teleological behaviourism, teleological explanation is regarded as basically externalist (though not of course evolutionary).

However tempting this version of teleological materialism seems on first sight, there is something basically wrong with it. It fails to account for how mental states *justify* anything. Why does having the belief that *p gives rise to q* and having the desire to achieve *q* justify doing *p*? Why shouldn't having the belief that *p gives rise to q* and having the desire to achieve *p* justify doing *q* for instance? The second structure is irrational; but in virtue of what is it irrational?

It cannot be argued that there is something about the nature of beliefs and desires that accounts for the rationality or irrationality of each structure. For the beliefs and desires are attributed only in virtue of being what would make the behaviour rational according to the first structure. One possible response would be that beliefs and desires cannot be consistently attributed according to the second structure of rationality. But the basic problem still remains. Why does it follow from the fact that a subject's behavioural tendencies can be consistently attributed in terms of some structure of belief/desire rationality that the subject's behaviour is thereby justified?

Millikan's teleological materialism to its credit does have an answer. Behaviour is rational when it leads to survival and

propogation of genes. As I have explained, I think this is an inappropriate answer; but at least it is some sort of answer. The internalist version of teleological materialism however accounts for rationality in terms of mental states and accounts for the attribution of those mental states in terms of rationality. It allows no other source for the basis of this rationality, and is consequently uselessly circular.

The accounts I have been criticizing in this section have been undiluted accounts. In each case someone may reply that their account has something to offer, but it needs to be mixed with something else. I have provided no arguments against any of the innumerable mixed strategies available for accounting for mental states. What I have tried to do is suggest the potency of a new pure stratgy - that of teleological behaviourism. In the rest of this thesis I aim to make this strategy more persuasive.

Chapter 2

Causal Explanation (1) - Causal Determination

2.1 Strategy

I will not pluck the details of my account of teleological explanation out of thin air. Rather, the account will follow inevitably from a proper account of causal explanation. Over the next two chapters I try to derive this account of causal explanation. The key is that causal understanding involves having some kind of theoretical model that represents the working of a real causal mechanism. There are many different things that can take on the role of this theoretical model: a scientific theory consisting of a set of laws and facts, a metaphor, a historical narrative, etc. All that is required is that the model *determines* a consistent set of *descriptions* of what should happen in certain circumstances. This will be a correct model, and thus constitute a correct explanation, if and only if it correctly models the way an actual mechanism works.

I will argue that what is *not* necessary is that the theoretical model embodies *universal* truth. I will argue that there is no objection to the descriptions it determines being false sometimes as long as this only happens when the model is applied to situations where the

mechanism it represents is not operating or is working but being interfered with.

Teleological explanation can be characterized by the sort of theoretical model involved. In chapter 4 I argue that in teleological explanation the theoretical model characteristically embodies practical rationality. The mechanism represented by such a model results in what should happen according to a notion of practical rationality. This seems to capture Aristotle's notion pretty exactly.

In chapter 5 I consider particularly externalist practical teleological explanation. Externalist teleological explanation depends essentially on an externalist version of practical rationality. Externalist rationality essentially involves means-ends analysis.

In chapter 6 I consider the implications of treating people as essentially teleological machines. I show how we can attribute beliefs and intentions. Then I can complete the task introduced in chapter 1 of analysing internalist teleological explanation in terms of externalist teleological explanation.

2.2 The Nature of Causal Explanation

One aspect of explanation is that it makes its object intelligible. So an explanation must involve a theory or rationalization which puts the phenomenon being explained in an intelligible context. Let us call this the theoretical aspect of an explanation.

If the explanation is to be the correct explanation, then it is not sufficient that it makes the phenomenon *understandable*. It must also make the phenomenon *understood*. So not just any theoretical rationalization will do. Now the question introduced in this section is whether the distinction between a correct explanation and an incorrect explanation can be accounted for entirely in terms of its theoretical aspect, or whether some other aspect needs to be considered.

On the one hand there is what I shall call the Humean view, according to which one may approach the ideal of making a phenomenon understood by making it more and more intelligible - i.e. by concentrating on the theoretical aspect. On this view the intrinsic quality of the rationalization of a phenomenon determines whether this rationalization constitutes a correct explanation.

On the other hand is the anti-Humean view, according to which the conclusion that a phenomenon should occur may be rationalized equally well by two competing explanations. On this view, the correctness of an explanation partially depends on whether or not that explanation represents the real thing that causally determines the phenomenon.

Consider an example. A piece of paper ignites; a lit match had been placed under it; also it was struck by lightning. One putative explanation of why the paper ignites involves the fact that a lit match is placed under the paper. The other involves the fact that the paper was struck by lightning. Let us assume that just one of the two explanations of the paper igniting is in fact correct; perhaps the

lightning got to the paper just before the heat from the match had time to work.

On the anti-Humean view, in terms of their theoretical aspects, both explanations may be equally good. Both result in the conclusion that the paper should ignite. Both make the phenomenon intelligible. So what is the source of the *correctness* of one explanation over the other? It must be that one of the two explanations describes the actual causal determination of the piece of paper igniting. Let us call this the material aspect of a correct explanation.

So the issue between the Humean and anti-Humean positions on causal explanation may be seen as the issue of whether or not a material aspect should be part of the account of explanation. Before considering the issue in this form, the theoretical/material distinction needs to be clarified.

An extreme version of the distinction is this. The theoretical component just concerns what makes something intelligible and is quite independent of the way the world actually is. Whereas the material component concerns the way the world is. According to this way of making the distinction, it is fairly trivial that the correctness of an explanation depends on some material component. Otherwise the correctness of an explanation could be determined a priori.

The triviality of this point suggests that the Humean/anti-Humean issue I am trying to consider does not depend on this way of making the distinction. Indeed, the Humean position I am trying to characterize is laid out in the following quotation from Rescher in a

way which explicitly requires both theoretical and material components understood in this way.

> *'What is it that differentiates an* actual *or correct explanation from one that is merely* possible *or conceivable?'*
> (Rescher 1970 p17)

> *'An* actual *explanation, therefore, must conform not only to the formal requirement that the explanatory premisses, if assumed as hypotheses, will render the explanatory conclusion relatively certain or probable; it must also satisfy the material requirement that the particular premisses be fact-asserting (true or highly probable) and that the general premisses be law-asserting, and represent generalizations which, being well confirmed, have earned the rubric of lawfulness.'*
> (Rescher 1970 p19)

An alternative way to make the theoretical/material distinction is this. The material component of an explanation concerns the *existence* of some real basis for the correctness of the explanation. It concerns the material existence of something that actually makes the phenomenon occur. The theoretical component concerns everything else: laws, facts about the way the world was before the phenomenon, etc.

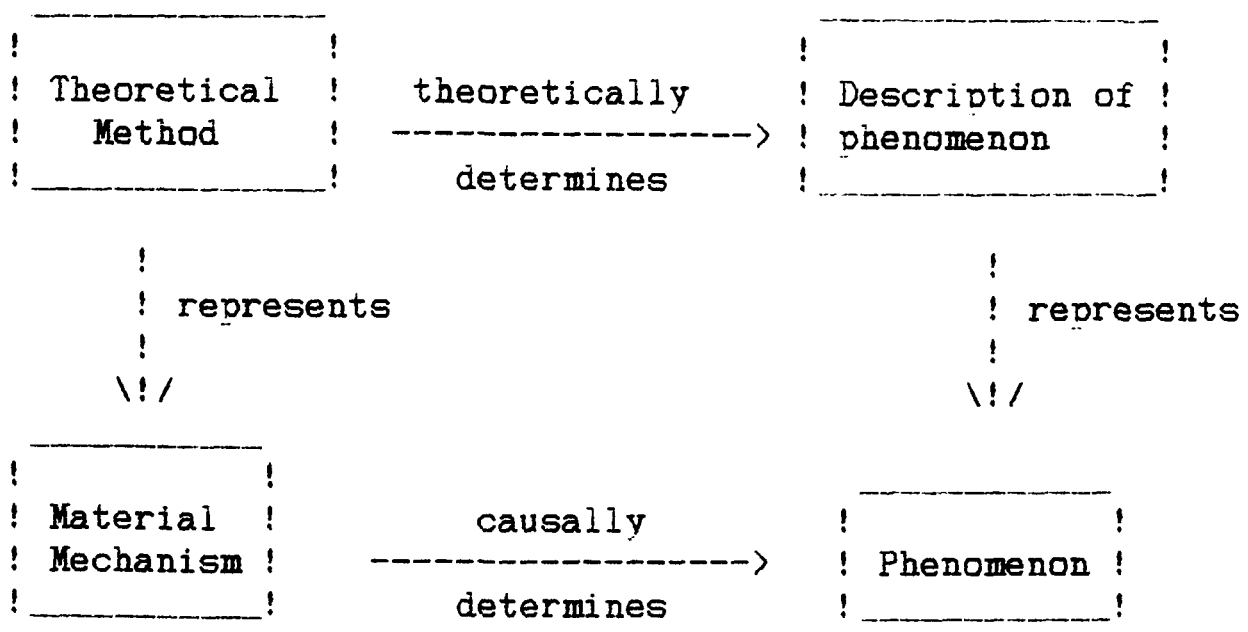
Given this way of making the distinction, it is not such a trivial matter that the correctness of an explanation depends on some material component. Rescher, in the quotation above, regards the consideration of a material component in this sense as irrelevant to the question of what differentiates a correct explanation from a merely possible explanation. His position is that by building up the theoretical side

of an explanation (i.e. the prior facts and the laws of nature) one can ensure a good explanation.

According to the anti-Humean view, a good explanation depends on both a theoretical and a material component. The material component must be somehow represented by the theoretical component. Corresponding to each component is a different sort of determination. The theoretical component determines not the phenomenon itself, but the *description* of the phenomenon. The material component determines the phenomenon itself. This point is made by Anscombe as follows.

> *'causality consists in the derivativeness of an effect from its causes. This is the core, the common feature, of causality in its various kinds. For example, everyone will grant that physical parenthood is a causal relation. Here the derivation is material, by fission. Now analysis in terms of necessity or universality does not tell us of this derivedness of the effect; rather it forgets about that. For the necessity will be that of laws of nature; through it we shall be able to derive knowledge of the effect from knowledge of the cause, or* vice-versa*; but that does not show us the cause as source of the effect. Causation then is not to be identified with necessitation.'*
> (Anscombe 1971 pp 7,8)

The following diagram - The Square of Explanation - expresses the anti-Humean account of explanation that is to be supplied. My claim is that no part of the square is redundant.



Let us apply the example of the igniting paper to this. I suggest that the top half of the square works just as well for theoretical determination in terms of the lit match as for theoretical determination in terms of the lightning. So there are two quite different ways in which the description of the paper igniting may be determined.

The ignition of the piece of paper is doubly intelligible, but the phenomenon itself is only to be explained in one of these two ways. Which way it is depends on the bottom half of the square. Each theoretical structure represents a mechanism - a lightning mechanism and a burning match mechanism. Only one of these causally determines the igniting of the piece of paper. Whichever theoretical structure represents the mechanism that causally determines the phenomenon (in this case the lightning theory) is the true explanation of that phenomenon.

On the Humean view the top half of the Square of Explanation does all the work. The main argument for this position is based on Hume's claim that we can never know about the existence of the thing that really determines a phenomenon: i.e. we have no epistemological access to the material aspect of explanation. (This claim can be neutral on the question of whether such an aspect actually exists independently of our epistemological access to it.) If we also assume that we do have epistemological access to whether or not an explanation is correct, it follows that the correctness of an explanation depends just on theoretical considerations.

A quick route to Humean scepticism about causal determination comes from yielding to a logicist confusion about what sorts of things we can know about; that is that all we can know about are propositions. If this is the case, then when we talk about the movement of the second hand of a watch we are really talking about propositions concerning the movement of that second hand. If we say that that movement is determined by the mechanism in the watch, we are really saying that the truths of some propositions concerning the movement of the second hand are determined by a theoretical consideration of the mechanism in the watch.

This is just the confusion of two types of determination that Anscombe is arguing against in the quotation above. To see that it is a confusion, let us assume that we can know about real causal determination and then show that this is quite different from knowledge of theoretical determination.

One way to see the difference is to put the movement of the second hand into the past (or the future). As long as the propositions we use are tenseless (i.e. 'the movement at time t_o is ... ', rather than 'the movement now is ... '), then what makes these propositions true is not linked to the same time as the movement itself. The movement was determined at a particular time before t_o, but how things are concerning propositions about the movement can be determined at any time.

Suppose that now is some time after t_o. How things are concerning the movement of the second hand at t_o *was* causally determined by something just before t_o. But how things are concerning a large conjunctive proposition about the movement of the second hand at t_o *is* logically determined by something else (a way of relating sentences to the world perhaps) either timelessly or at least after t_o.

The movement of the second hand and propositions or descriptions about the movement of the second hand are determined by totally different things. The movement itself, being something that really happens at a particular time, must be determined by something concrete - the watch mechanism for example. Whereas a proposition or description about the movement of the second hand is determined logically by a theoretical technique.

One excuse for failing to make the distinction properly is a looseness in our use of language in sentences like, 'The motion of the apple is determined by Newton's Law of Gravity.' Now the motion of the apple is not in fact determined by Newton's Law of Gravity; it is

determined by the mechanism of gravity. What is determined by Newton's Law of Gravity is the *description* of the apple falling.

2.3 The Need for Processes

Proponents of the Humean view may complain that they never confused theoretical and material determination. They may feel that they have an independent argument to show that we cannot have knowledge of real causal determination.

> *'When we look about us towards external objects and consider the operation of causes, we are never able, in a single instance, to discover any power or necessary connexion, which binds the effect to the cause, and renders the one an infallible consequence of the other.'*
> (Hume's Enquiry §vii part I)

What are the assumptions behind this argument? Central to Hume's ontology is what might be called the snapshot approach. Impressions are temporally located flashes of experience. Sense-data likewise are to be understood as cross-sections of experience at particular times. And if Hume can go so far as to identify a world outside his experience, his picture of the world must be thought of in the same way as being built from snapshots of the world.

For Hume, a state of affairs is something that one can identify by building up occurrent and potentially occurrent observational properties of a snapshot in time. However closely one looks into this snapshot,

one will never see the causal connection with another snapshot. Even by examining in the minutest detail an infinite continuous series of such snapshots one can never observe any determination. The future state of affairs does not appear preordained somehow anywhere in the series of states leading up to it. One could imagine the series of states leading up to an apparently determined state all happening, just as they do in fact, but the final state not happening. So what determined the final state cannot be seen after all.

If we find this argument convincing, we do not have to use it to deny real causal determination. It can be used instead to deny the snapshot approach to epistemology. We could turn the argument upside down, taking as an assumption that there is real causal determination and using Hume's argument as a reductio ad absurdum of his epistemological assumption. If we could identify processes as entities irreducible to snapshots, then we would have no difficulty in identifying real causal determination.

Having events in one's ontology does not necessarily save one from Hume's argument. The snapshot approach can be applied to event ontology, so that the only events that one can basically identify are instantaneous changes of state. If one's world picture must be built up out of distinct snapshots, whether the snapshots are static states of affairs or instantaneous changes of state, one can have no place in that world picture for real causal determination.

Although this argument should make us look hard at a way of identifying things not based on the snapshot approach, it is not an

obviously contradictory policy simply to reject causal determination instead. However there is a Kantian argument that, if valid, makes it extremely pressing that we find a way of identifying determinate processes in addition to snapshots.

In the Second Analogy from The Critique of Pure Reason Kant writes as follows.

> *'the* objective relation *of appearances that follow upon one another is not to be determined through mere perception. In order that this relation be known as determined, the relation between the two states must be so thought that it is thereby determined as necessary which of them must be placed before, and which of them after, and that they cannot be placed in the reverse relation. But the concept that carries with it a necessity of synthetic unity can only be a pure concept that lies in the understanding, not in perception; and in this case it is a concept of the* relation of cause and effect, *the former of which determined the latter in time, as its consequence [Folge = follower] not as in a sequence that may occur solely in the imagination.'*
> (Kant's Critique of Pure Reason B234)

I paraphrase Kant's argument as follows. We could not see static states of affairs or instantaneous events as ordered in time objectively unless we saw them as part of objectively real processes: i.e. unless we could identify such processes in the world. Looking at the world as structured into processes provides a non-arbitrary order to our snapshots of the world which cannot be supplied if we can only identify the snapshots and not the processes.

The following rejoinder might be made. We can assume that most continuous properties change continuously. So fitting together

snapshots is like putting together a jigsaw puzzle. There is only one way to do it that maintains a smooth transition between each piece.

The Kantian reply is that without being able to identify processes there can be no objective basis to the assumption that properties change continuously over time. The proponent of the rejoinder may then be forced to say that the jigsaw puzzle solution actually constitutes what is the correct time sequence. If there is no good jigsaw solution, there is no time sequence in reality.

But if the proponent of the rejoinder does have to come to this, they have thereby given up any chance of being able to identify an *objective* time sequence. Suppose that there were two orderings of snapshots, each equally good as far as the jigsaw puzzle solution goes. Not only could we not tell which ordering was correct, but also since the jigsaw puzzle solution is supposed to provide the actual criteria of what is correct, both orders *are* equally correct. There is nothing in the world determining the correct time sequence over and above this.

When whatever determines whether something is true coincides with whatever determines whether we say it is true, then objective truth is lost. If the jigsaw puzzle solution to how we determine a time sequence of snapshots is in fact what determines the correct time sequence, then the correct time sequence is not an objective matter.

Of course, it has been argued by some snapshot ontologists that time is not real after all. But as with Hume's argument for the unreality of causal determination, it is more attractive, having

accepted the validity of the argument, to use it as a reductio ad absurdum on the epistemological assumptions.

The snapshot ontologist may try to avoid this slide into scepticism by claiming that natural and artificial clocks provide an objective basis for a time sequence. But this is grasping at straws. A clock only serves this purpose if we can identify the process of a clock working. Otherwise there is no reason (other than the jigsaw puzzle solution) to place a snapshot showing the clock at 12.00 p.m. just before a snapshot showing the clock at 12.01 p.m.. Having many clocks all agreeing in the time order they show may explain why we feel compelled to fix a particular order on our snapshots, but it does not provide an objective basis for that order.

In presenting these two transcendental arguments for the identification of processes I seem to have presumed an exhaustive dichotomy between snapshots and processes. A natural objection to this is that there may be categories which require more than the identification of snapshots, but do not obviously require the identification of processes. An example of such a category is that of persisting objects.

Some people think that persisting objects are processes. But this is not my line of defence against the objection. I take it to be essential to a persisting object that something about that object remains unchanged if it is left to itself. At any rate, it is not essential that something changes. However, change is an essential

aspect of processes. If a process is left to itself, things will change.

This is very rough and ready, and it will be worked on later. But notice that it is this essential involvement with change that is required by the two arguments of this section. Without being able to identify things which essentially involve change we would not be able to identify real causal determination. Also, without being able to identify things which essentially involve change we would not be able to identify real temporal order. The snapshot ontology is just a particularly vivid example of one that does not include things which essentially involve change.

There is another requirement in addition to this one which is imposed by these arguments. That is that we must be able to identify things which involve *determined* change. This can be illustrated by contrasting determinate processes with events. Both these things essentially involve change. But with a determinate process a change may be said to be determined by that process just in virtue of belonging to it. (This should be taken as a definition of 'determinate'. From now on I will always mean determinate processes when I talk of processes.) For example, fire spreading is a determinate process. If the change of a certain piece of wood from non-ignited to ignited belongs to a particular process of fire spreading, then it follows that that change is determined by that process. This is not generally true of *events*. Consider the event of the final game in the 1990 FIFA World Cup. No part of that event was determined by it.

The identification of such an *event* does not help to account for how we identify real causal determination. But if we have a way of identifying a *determinate process* and can say that some change belongs to that process, then we can, in virtue of that, identify real causal determination. If A is the change from the state where the process is not happening to the state where it is happening and B is some change that belongs to that process, we can thereby say that A causes B.

2.4 The Nature of Processes

By identifying a need for determinate processes I have not shown conclusively that there are such things. To make further progress I will adopt a mixed strategy. I assume initially that we can identify processes and then pin down their nature with constraints from three different angles.

1. What notion of process does one need to avoid the sceptical implications of the snapshot ontology considered in the previous section?
2. What does the nature of our language about processes dictate about their nature?
3. What constraints on the nature of processs does the possibility of identifying them impose?

If I am left with a plausible notion of processes satisfying these three constraints, I will take it as a pretty good argument that we can

identify entities corresponding to this notion. Once we are convinced that there is available an identifiable material aspect to explanations, it is a small matter to become convinced that the distinction between a correct and incorrect explanation should depend on this.

I shall start by considering a fairly bad account of processes - the reductionist account.

The assumption underlying a reductionist account of processes is that a process cannot be identified except as a collection of events. Events are regarded as constitutive parts of a process in the sense that the process is regarded as nothing other than those events put together. For the sake of argument at this stage, I will just assume that events are just changes of state: the objections to this account do not depend on what notion of events we are working with.

On the face of it this reductionist account has some plausibility. The event of my hair turning white is part of the process of my growing old. If you put all such parts together you would seem to get the whole process of my growing old. On this account, the process of my growing old just is the complete series of such events.

As it stands, this could only be the starting point of an account. For, not just any collection of events may constitute a process. A typical approach now would be to formulate a set of necessary and sufficient conditions for such a collection to be a process.

For example, perhaps the change between contiguous events in a process must be largely continuous (continuous for the majority of their properties). Indeed one must have some such condition to rule

out nonsense collections like that consisting of my putting my left foot forwards one hundred times (when in fact I was just walking along).

On the other hand, this may be too strong a condition. Playing a game should count as a process, yet there may be gaps between the moves.

Another condition is that of causal connectedness. For example, causally unrelated moves cannot make up a determinate process of playing a game. Or again, if a ball is rolling along a table, picked up and rolled back, we have a continuous sequence of events concerning the ball. But what is happening to the ball at first is not the same as what is happening at the end. So there is not just one process happening here. An extra condition is required to delimit processes, and causal connectedness seems to be it.

Also, without this condition we could not ensure that a series of events constituted a *determinate* process. There would be no conceptual connection between an event being part of a process and the event being determined by the process.

Unfortunately however, by including causal connectedness in the notion of a process, the analytical point of talking about processes is lost. For, the strategy here was to show how causal connectedness can be identified in terms of the identification of processes.

Moreover, the whole point of saying that a process determines an event is lost on this conception. What is the point in saying that a particular process determines a particular event if that process cannot

be referred to except as including that event? On this account, once we know what a particular determinate process is, there is nothing to be added by saying that the process determines some event.

Consider the series of the events of a ball passing over point A_1 at time T_1, point A_2 at time T_2, ... , point A_n at time T_n. Assume that any other required conditions such as continuity and causal connectedness are satisfied. Now the ball passing over point A_n at time T_n is not in any way the result of that series of events. It is the result of the previous event in the series (and, because of causal transitivity, of all the previous events). But we explain nothing by saying that the event is the result of the entire series of events *including itself*. If we added to the end of the series the event of the ball turning purple at time T_{n+1}, this would not make that event the result of the augmented series, even if its turning purple was caused by earlier members of the series.

Consider on the other hand the ball rolling along an inclined groove from A_1 to A_n. This is a process: the event of the ball passing over point A_n at time T_n is part of that process and it is the result of that process. This is because the process of the ball rolling along the groove has an identity beyond the identity of its constituent events. It is identified with reference to a physical system - a ball and inclined groove. Identifying the process as the working of this system is done quite independently of identification of the event of the ball passing over the point in question.

We perceive the need to identify processes in order to talk sensibly of determination by processes. But having gone that far we cannot stop with the position that a process is identified as a series of events. If a process is to be capable of determining anything, it must have independent reality.

The impossibility of maintaining this reductionist position can be shown by considering the requirements that derive from the need to identify and reidentify a process. The point that will be argued is that if it is possible to identify and reidentify a process, then a process cannot be a collection of events.

If a particular process is identified with a particular collection of events, then it could not be any other collection of events. A *collection* of events could not consist of different events; its identity is dependent on the identity of each of its constituent events. Yet the same process could involve different events. Therefore a process is not a collection of events.

The assumption that a process could involve different events needs to be examined. It is only significant if the events that could be exchanged while the process still happens are those that the process is supposed to be constituted of. This is in fact the case. For a process may be identified before all its constituent events have occurred. The later events are determined by the process unless something interferes. But what if something does interfere? Does this mean that the process was not happening after all? No, unlike an

event, which is primarily identified only after it has happened, a *process* is primarily identified while it is happening.

There is no escape from this argument in the claim that a process is only to be identified as the collection of events occurring in the very narrow space about the time of the utterance. For a process may continue for a while: and the same thing is happening after a while as was happening before. If a process is identified with a collection of events, then that collection must last from the beginning to the end of the process.

One consequence of identifying a process with a collection of events is that whether a certain process is happening now is not established until the last event of the process has happened in the future. What is happening now (i.e. the process) must then logically depend on what happens in the future. So present tense process identification must be regarded as a disguised form of future tense prediction.

This is not acceptable. We must have an account that allows that what is happening now might end in several different ways depending on circumstances. This shows up the difference between processes and events. The process of a ball rolling across a table may continue until the ball falls off the edge, or it may be interrupted. But it makes no sense to say that the event of a ball being rolled across the table, picked up and rolled back could have been stopped in the middle. If the ball had not been picked up and rolled back, then it would not have been *that* event that happened.

The reductionist view which ties a particular process to a particular set of events is too restrictive. What about a less restrictive view which takes processes to be essentially *structures* of events, where the identity of the whole structure does not depend on the identities of the individual events? On this view, one thing that is essential to a particular process is that its constitutive events satisfy some structural conditions linking them together; but these conditions are not so tight that the same process might not have different events as part of it.

I take it as a natural starting point that it is essential to the identity of a particular process that its constituent events do satisfy some structural principle. I want to consider two accounts that take off from this starting point.

One account regards this structural constraint on constituent events to be *all* that is required to identify a process. According to this account, one can tell that a particular process is happening by telling that events that have been happening are part of the right sort of structure of events. It follows from this that at any one time only part of a particular process is present. When the process is re-encountered what is really happening is that another part of the process is identified as belonging to the same process as the first part that was identified. So processes are identified and re-identified in quite a different way to the way objects are identified and re-identified.

Evans seems to have just such an account of processes in mind in the following quotation.

> *'If the concept of reidentification is to be used in connection with processes, it must be understood that it is being used in a different sense from that which it has in connnecton with things. We reidentify a process when we hold that an occurrence encountered at one time is* part of *the same process encountered at another, but it is a distinctive (and some have thought incoherent) feature of our conceptual scheme of material bodies that we suppose an object to be both present* as a whole *on one occasion, and literally identical with an object present as a whole on another.'*
> (Evans 1985 pp257,258)

The other account I want to consider regards a process as a continuant - i.e. as something that can be identified as a whole at one time, and then at a later time the *very same thing* can be re-identified as present. What characterizes this account is that the essential nature of a process is identifiably present at all times while that process is happening. So, there must be some essential nature to a process, in addition to the structural constraints on constituent events, through which the process is identified. For the structural constraints are not identifiably present at all times that the process is happening. On this account, the way to identify a process is to identify its *continually present* essential nature.

Aristotle's notion of processes discussed in *Physics 3* is much like what I am after here.

> *'Def. The fulfilment of what exists potentially [a capacity], in so far as it exists potentially, is motion [a process]. ... Examples will elucidate this definition of motion. When the buildable, in so far as it is just that, is*

fully real, it is being built, and this is building. Similarly learning, rolling, leaping, ripening, ageing.'
(Aristotle 201a10-18)

According to Aristotle, the capacity for being built is the underlying nature of the process of building. When it is realized, building is going on. When the building has finished there is no such capacity any more. So a process in Aristotle's sense has complete existence from the time the building starts to the time it is over.

One might object to the account of processes as non-continuants that it fails to distinguish processes from structured events. But perhaps this in itself is no objection. The response might be that when one identifies a structured event at a time between the first and last of its basic events, then one calls it a process.

Consider something that should count as a process according to this view - the structured event of the water level on a tank rising from one mark to another. A characteristic structure of events is in train involving continuous changes of properties and causal connectedness between changes. All that I assume that is lacking is some continuously present ground to these changes.

Note that this structured event is not the process of *the tank being filled up*. Part of the process of the tank being filled up is some agency or power responsible for the structure of changes. I am assuming that this is not part of the structured event of the water level rising from one mark to another.

Why might we not identify a process of the water level rising characterized by a structure of events without requiring the existence of any other underlying nature to the process? A process identified later as that of the water level rising could be identified with the earlier process if and only if the events of the water rising that characterize the process initially continued uninterruptedly (and causally connectedly) until they identified the process later.

As a matter of fact, we do not speak of the water level rising (understood in this way as an identifiable particular) in the way we do of something like fire spreading (understood as a determinate process). We may be interested in knowing whether the some *sort* of thing is happening now as was happening earlier. But why should we be interested in whether a *particular* entity identified now is the same as an entity identified earlier? We are not interested in whether or not a series of events was interrupted for an infinitessimal moment for its own sake. We are only interested in this in as much as it reveals a change in the substantial, relatively abiding ground underlying the characteristic events. And the sort of thing we are considering like the water level rising is not identified with respect to such a ground.

This does not prove that we cannot have entities of this sort. It just indicates that there is no point in having them. What makes such putative entities even more ridiculous is that we would hardly ever have any basis on which to reidentify one. A putative entity of the kind we are considering, like the rising of the water level, is supposed not to be identified as essentially a persisting thing. Identifying

such a thing does not provide reason to suppose that it will last even a microsecond longer.

So, having identified such a thing, it may stop for no reason and then be replaced by the same sort of thing. (Assume some deviant causal chain to keep the causal connectedness condition satisfied.) The series of events may be interrupted for no reason. We could have no reason for denying that this has happened in any particular case unless we had been keeping a watch on the series of events continuously throughout the time.

This continuous observation could be done through a surrogate like an alarm system. But it still limits one's ability to identify such putative entities drastically, and shows that at the very least it is highly impractical to have such entities in one's ontology.

A more serious problem is that such entities could only ever be identified retrospectively. We could never have any conclusive reason to say that the water level is rising, only that it has been rising. If the identification criteria just pick out the series of events up to the point of identification, and nothing inherently persisting is picked out, then they do not pick out anything that is persisting across the present moment. Yet when a process happens it happens across that moment.

Suppose that a process is occurring until time T and at T it is no longer happening. If we are identifying the process as having as its essential nature a characteristic series of events, then that process is

correctly identified at T. But what is identified is not a process that is happening, but a process that has been happening.

If the essential nature of a process is identifiably present while the process is going on, then it is possible to establish that the process is going on. If there is no such continuously present nature, then it is not possible to establish that the process is going on: it is only possible to guess.

The fact that processes are primarily identified while they are happening and structured events are primarily identified after they have happened is not just a matter of definition. (I.e. it is not just a matter of saying, 'Call this a process if you are in the middle of it, otherwise call it an event.') If I ask, 'What is happening here?' I usually want to know what process is going on. The answer, 'Wait and see,' is not merely frustrating: it is entirely inappropriate.

The great advantage of the continuant view of processes is that, according to it, there is a way of telling that a certain process is happening which can be applied from the moment the process starts to the moment it finishes. That way is to identify the presence of the underlying essential nature. This is not possible with the identification of structured events.

This view also fits in better with the way we talk about processes. A process is something that is happening now. Whereas a structured event is something that occurs around this time. If a process were a structured event, then it would be more appropriate to say that what was happening right now, at this precise moment, was just one of the

constituent events. But we feel no inclination to talk like this (unless of course we are in the grip of a philosophical theory). We say that what is happening at this precise moment is an apple is falling, a fire is spreading, a child is growing, etc. We refer to the process as a whole and not just to a constituent event that occurs at the time of utterance.

I do not claim yet that the case has been entirely proven in favour of the continuant view of processes. But in the next section I will assume that this view is correct and investigate how such processes might be identified. I will show that the identification of continuant processes is both possible and extremely useful.

2.5 The Identification of Processes

There is an apparent problem with this account of processes as continuants. It concerns what to do about the structural constraints on the series of constituent events. The continuant view that I have been proposing takes the presence of a particular underlying nature to be not only essential but also *sufficient* for the identification of the corresponding process. But, given this, it seems to be at least logically possible that this essential nature be present while the structural constraints on the ensuing events are not satisfied: the wrong events or no events at all occur. This possibility contradicts our starting point that it is essential to the identity of a process

that the right sort of events (i.e. events satisfying the structural constraints) occur.

There are three possible responses. One is to say that the characteristic structure of events of a particular process is *nominally* but not *really* essential to the process. For example, suppose that the underlying nature of a water heating process were present but the water did not heat. According to this response, we should say that some process is happening which *would be* water heating if the temperature of the water started to go up, but cannot be so described since no such changes occur.

The trouble is that this response only seems to be available in certain special circumstances. Suppose that I was carefully pouring ice cold water into the top of the tank so that the effect of the heating element was being exactly cancelled out. Here it might be appropriate to say that the heating process was happening, it was just not resulting in its characteristic effects. But if no such interference is occurring and the events do not materialize, it seems quite wrong to say that the process *however described* is happening.

It is part of the *real* essence of a process that it involves particular kinds of change. The only allowable weakening of this condition would be to say that the real essence of a process is that it involves *potential* for a particular kind of change - i.e. such change occurs if nothing interferes. So we have not eliminated the problem of conflicting real essences with this response.

A second response is to regard the underlying essential nature of a process as necessary but *not* sufficient for the identification of a process. So the real essence of a process could be a conjunction of the underlying nature and the characteristic structure of changes. The process of the water heating up would not be happening unless the structure of temperature changes occurred and the underlying heating mechanism were present.

Unfortunately, this response involves sacrificing some of the most attractive features of the continuant view. In particular, we would not have a way of telling (without having to guess the future) that a process was happening rather than just that it *had been* happening. Another feature of the continuant account is that, having identified the process, one has identified what determines the characteristic events. This feature also is lost if we accept this response. For, according to the response, a process is not to be identified as the process it is except in virtue of it involving the characteristic events it does. There is not the logical separation of a process and its effect that is essential for causal determination to make sense.

There is a third response that may avoid the problems of the other two. This is to accept that the presence of the underlying nature is sufficient for the process to be happening and to accept that the characteristic events must occur when that process is happening and nothing interferes, but to deny the possibility that the underlying nature be present while the characteristic events do not occur and nothing is interfering. This may be done in the following way.

Suppose that we want to identify a process characterized by a certain type of event structure. We do this by picking out an underlying structure that is common to the duration of processes so characterized, and where, on those occasions where it is present but the characteristic events do not materialize, we can always locate some interfering factor. If we cannot do this, then we cannot identify a determinate process characterized by those events.

We have only successfully picked out the underlying nature of a process if what we have picked out *always* involves the characteristic events when nothing interferes. What is the status of the word 'always' in this claim, and how can one know that the characteristic events will always follow?

Imagine that the underlying nature of a process is apparently identified on many occasions. Then one day the identification is made, nothing interferes, but the characteristic events do not occur. What we identify on this occasion is not the underlying nature of the process, for the process is not occurring. If it is the same thing that we identified on the previous occasions, then on those previous occasions what was identified was not the underlying nature of the process either.

So the necessity that the results follow if nothing interferes must be regarded as of the strongest sort. No exceptions are allowed. If what seemed to be a process failed once (for no reason) in the history of the universe, then we would have misidentified that process on all other occasions. Of course, if this did happen, we would probably be

able to salvage something through adjusting the characteristic events of the process or the conception of its underlying nature.

There is no epistemological problem with this, for one does not need to know with absolute certainty and accuracy the underlying nature of a process in order to be able to identify it correctly. One may have good reason to suppose that there is an entity there to be identified if the alternative of randomness is statistically unlikely. The nature of the process is identified through feeling one's way around it.

Experiment and experience give one an idea of the shape of an entity, so that one can successfully identify it in the future. But in order to do this, there is no need to have absolute certainty in one's identification. So the element of necessity can be identified and yet one have no certain knowledge of the future (even conditional on nothing interfering). One can identify correctly the essential nature of a process, which when correctly identified will always be associated with certain kinds of future event unless anything interferes, without thereby knowing with absolute certainty that those events will occur unless anything interferes (i.e. without thereby knowing that it *is* the essential nature).

I can identify correctly a cup in front of me, even though I do not have absolutely certain knowledge that I am not having a hallucination in which case the characteristic signs may disappear for no reason. If I have correctly identified the cup, then I have identified some natural necessity; for what I have identified cannot just disappear for no

reason, but must continue to give those characteristic signs unless anything interferes. Such natural necessity should not be confused with predictive certainty. For it is always possible that there is no cup there and my identification has been wrong.

As with objects, so it is with other entities like processes. The possibility of doubt - the possibility that one's predictions of the materialization of the characteristic signs (events or whatever) may turn out false without there being anything interfering with the entity - is the possibility that one's identification has failed. It should not be thought of as the possibility that one's identification has succeeded while the world has failed instead. For, if one's identification is successful, then the world cannot just fail to deliver the goods.

There must be something it is for a mechanism to be working which is identifiable before the results (the characteristic events) happen, yet which determines these results. If the mechanism is working, the results will happen unless something goes wrong. This is what gives causal determination the feeling of necessitation. The fact can be expressed as a law-like assertion with a logical status.

The form of the logical law is this:

> *If there is a determinate process to be identified, then if it is correctly identified, the characteristic events must occur unless anything goes wrong.*

It is a law with plenty of slack. Suppose that one has a concept of a process from paradigms characterized by a type of event structure, and one apparently identifies the process through identification of a

mechanism, but the characteristic events do not occur. There are four possible ways this may happen.

1. One has identified a process, but what one took to be a characteristic event of that process is not. (This trick can only work for an exceptional event. If too many events turned out to fail, then we should have to say that our process identification had somehow failed.)

2. There is no process to be identified here. (This option is only to be taken if all else fails.)

3. What one took to be criteria for the identification of the process were not. (The criteria can be altered to identify more accurately the process.)

4. The process was correctly identified, but something went wrong. (Some other process can be identified, one of the characteristics of which is that it makes the first sort of process go wrong.)

So the law is as much an expression of a method of settling in to a way of identifying processes as it is a fact about the world. It's truth does not constrain the world at all because of the catchall possibility 2. Still it is what accounts for necessitation in nature. For once a method of identifying a process has settled in - i.e. there is an abiding structure identified in way W, characterized by event structure of type E and interrelating with all other processes according to their characteristic events - then if W is satisfied E must happen as long as none of the potentially interrelating processes interferes.

Methods of identifying processes are not just imposed on the world. We could never account for natural necessitation this way. They settle in with feedback from the world and from other process identifying

methods forcing alteration or abandonment until they fit. But once they do fit, the world is determined according to them.

I am not assuming the uniformity of nature in making this claim. If there was not uniformity of nature, no method of identifying processes involving abiding underlying structures would fit. There would be no processes.

It may be thought that this account has swung back towards a Humean account. If one substitutes my talk of methods of identifying processes with talk of laws instead, one has a Humean account of explanation. But there is a difference and it is that, according to my account, the settling-in method fixes what entities there are; whereas, on the Humean account, it has entirely epistemological significance.

Let us apply this account of process identification to the process of fire spreading. The characteristic type of structure of events involves unburnt material adjacent to burning material igniting. The process itself is identified as the proper working of the fire-spreading mechanism. This involves a continuous expanse of inflammable material with plenty of oxygen and some of the material on fire. There may be more conditions to the identification of this mechanism, but not an infinite number.

If these conditions are met, the future occurrence of the characteristic events is determined. Suppose that the characteristic events do not occur (i.e. the fire stops spreading). Either the process was misidentified after all; one's notion of the essential nature of fire spreading should then be refined. Or something else happens that

stops the process working. A thunderstorm douses the flames perhaps. In this case, one can identify another process that intersects with fire spreading in this characteristic way.

The way I have accounted for mechanisms satisfies Aristotle's condition that there should be some nature underlying a process that is both definitionally linked to the characteristic changes of that process and at the same time a real nature. Aristotle had the same ontologically robust notion of processes as the notion I am trying to justify. He defined a process as the fulfilment of a certain potentiality in so far as it exists potentially. So, if there is an essential nature that can be defined as the capacity for a certain type of event sequence, the fulfilment of that nature is the process. These capacities are at the same time real aspects of the subject and definitionally tied to being capacities for certain changes.

> *'It is the fulfilment of what is potential when it is already fully real and operates not as* itself *but as* movable, *that is motion. What I mean by 'as' is this: Bronze is potentially a statue. But it is not the fulfilment of bronze as* bronze *which is motion. For 'to be bronze' and 'to be a certain potentiality' are not the same. If they were identical without qualification, i.e. in* definition, *the fulfilment of bronze as bronze* would *have been motion.'*
> (Aristotle Physics 201a28-34)

There is a possible source of confusion that needs to be dealt with here concerning what fulfilment or actualization is supposed to be. If we say of a property - say redness - that it is actualized in a certain subject at a certain time, then that just means that the subject has

that property at that time. But when we say of a capacity or disposition - say the capacity to melt - that it is actualized in a subject at a certain time, there is ambiguity. We might just mean that the subject has that capacity; i.e. that the subject is in a state such that if certain conditions are met (e.g. temperature), then it will melt. Or we might mean by the actualization of the capacity the meeting of these conditions; i.e. that the substance is in fact melting.

It seems natural to read Aristotle as making the distinction in the second way. According to this reading, a capacity may be possessed by a substance without being actualized. But this reading sets up the temptation to a quite spurious way of thinking about processes. For it suggests that the essential nature of a process may be a possessed but not necessarily actualized capacity. Then the actualization of this capacity - i.e. the meeting of its conditions - is the fulfilment of the already existing fundamental nature of the process. Unless these conditions are met, the essential nature of the process will remain dormant, present but not active.

However the essential nature of a process should not be thought of as the capacity, but as provided by the capacity and actualization together. For, the very same nature which actualized in one way would identify one process might identify a different process when actualized in another way.

Moreover, two quite separate capacities might, when actualized, identify the very same process.. Suppose that all the conditions for the process of fire spreading are satisfied except the presence of an

initial flame. We might say that an unactualized capacity for the spread of fire was present. But then suppose that an initial flame is present and all the conditions for the same process are present except the presence of Oxygen. This is another unactualized capacity for the spread of fire. Neither can count as the underlying nature of the process. The underlying nature must be an actualized capacity.

While one has only identified an unactualized capacity, one has not identified the essential nature of a process. A nature can only be described as a dormant capacity in virtue of its sharing all but a few conditions being met with an actualized capacity which is the real essence of some process. The underlying nature of a process is not something that needs to be switched on for the process to be happening. If the underlying nature is present, nothing else is required; the process will be happening.

So the underlying nature of a process seems to be naturally divisible into two parts: a capacity and conditions for the actualization of that capacity. In more up-to-date terminology, we might divide the underlying nature into a mechanism and operational conditions for that mechanism. As the example earlier showed, there is no hard and fast way of making this distinction. How to divide the underlying nature into mechanism and operational conditions is at least partly interest-relative. The rough idea is that the operational conditions concern the more fluid aspects of the underlying nature. But I suspect that nothing much depends on this distinction.

When a process involves a structure of different changes through time, the underlying nature must change through time to account for this. However, something about the underlying nature must remain the same, since the process remains the same. So the operational conditions may evolve as part of a process while always satisfying some structural principle.

How does interference work given this notion of a process? One way is when there is direct interference with the characteristic events of a process. This is the case of the intrusion of ice cold water into the water heating process discussed earlier. Here, I think we would say that there are two conflicting processes happening. It no longer makes sense to describe what is happening as the process of the water heating. But it may make sense to say that the same thing is happening as would be happening if the water were heating, only the description of it as water heating is now wrong.

Another form of interference is with the underlying nature of a process. When this happens, the process stops happening. This is the case whether the interference is with the mechanism or with the operational conditions. But there is an interesting special case where the interference is with the operational conditions. It is sometimes possible to describe what continues to happen as the process of the *faulty working* of the mechanism. For example, imagine an automatic production line wrapping a chocolate bar. Halfway through the process I push the bar out of alignment. After this the machinery continues to operate on the bar, but it is no longer wrapping it; instead it is

generally making a mess out of it. What is happening is that the mechanism is still working, but the process is quite different.

When nothing depends on the distinction between proper and faulty working of a mechanism, I will generally talk of the working of a mechanism to mean the *proper* working of the mechanism. It should be remembered that the basic notion here is that of proper working. This corresponds to the characteristic events of a process.

2.6 Identification of Causal Determination

Given the continuant account of processes, it becomes a non-trivial question whether a particular event belongs to a particular process (identified through a particular mechanism). The fact that an event belongs to the characteristic type of structure of events associated with a certain process does not necessarily mean that it is part of that process.

So how do we tell that an event belongs to a certain process? We identify the underlying mechanism; we see if anything is interfering; if not, we look for the characteristic events; if this particular event has no rivals, then it must be determined by that mechanism and belong to that process. We can go in the other direction and look for rival processes to which the event may belong. If there are no candidate processes other than the one we are interested in, then the event

belongs to that process. If there are rivals to the process or to the event, then we must match off rival processes and events.

A similar analysis applies to the question of how a process is reidentified over time. Suppose one identifies fire spreading at one time and then again an hour later. One can simply infer that one has identified the same particular process from the fact that nothing went wrong with the first process and there is no other contender for a process now/with the process earlier or indeed for a process earlier to be identified with the process now. If one can assume that processes continue unless something interferes or goes wrong, then process reidentification is seen to be possible.

(This assumption, that the underlying nature of a process must be an abiding thing, is necessary for the re-identification of processes to be possible. It is not something I have discussed here, but it seems likely that it *would* have to form part of a complete analysis of the nature of processes.)

According to our method of identifying the process of fire spreading, we can provide criteria for reidentifying any particular such process over time. If the fire-spreading mechanism is uninterruptedly present and working and uninterfered with over time, then the same process is still occurring. A particular later event is identified as belonging to this original process if in addition to knowing that the process is still continuing, one can match off processes and characteristic events, and this event can be seen to belong to no other process than this one.

So if I come back to a fire an hour later and see a stick burst into flames, I can identify this as belonging to the original fire if I know that there is no other fire spreading process (or any type of process that has igniting sticks as one of its characteristic events) that this event could belong to. If there are several fires spreading simultaneously, then I have to find a way of tracing through the particular events belonging to each process. When no such way is even theoretically possible, then we must say that different processes have become inextricably jumbled and they have lost their individual identities. This would happen for instance if several fires joined up.

I suspect that this account is rather simplistic. The aim is just to sketch how the identification of an event as being part of a process might be done. In the same spirit of sketching out an account, it is possible to extend this to the identification of causal determination and of causation itself. An account of the notion of causal determination follows naturally from the notion of a determinate process. Thus:

> *We say that a mechanism causally determines an event if and only if that event belongs to a determinate process identified as the working of that mechanism.*

We have seen how a determinate process is identified and we have seen how a particular event is identified as belonging to a particular process. So now we can see how real causal determination is identified. Note, according to this account, there is no need for the determined event to occur at the end of the process. So long as it

belongs to the process at any stage, then it is causally determined by the mechanism.

Next, we can define immediate causation.

> *An event A immediately causes an event B if and only if A is identifiable as a change from a state in which a process resulting in B is not happening to a state in which such a process is happening.*

So we may say that the event of something being ignited at place P causes something to burst into flames at place Q if and only if the latter can be identified as belonging to the same process as is identified by the state of the fire having been ignited at P.

Non-immediate causation is analysed recursively as a chain of immediate causation. If immediate causation required absolute contiguity of cause and effect, there might be a theoretical problem in getting causation even of a non-immediate sort to work over distances of time and space. However, there is no such problem, since on this account immediate causation itself can involve large leaps through time and space.

> *Event A causes event B if and only if either A immediately causes B or there is another event C such that A causes C and C immediately causes B.*

Chapter 3

Causal Explanation (2) - Intelligibilty

3.1 Understanding

In this chapter I provide an account of what it is for some fact R to be a reason why some phenomenon P occurs. When a phenomenon is explained by the provision of a reason some kind of understanding is achieved. This understanding may be partial in the sense that one understands why P occurs without knowing the answers to all the possible questions about it. But this is consistent with there being something here which is completely understood.

There may be several things involved in understanding why a particular phenomenon occurs. My interest here is only in that kind of understanding that is associated with *reasons* being given to explain why the phenomenon occurs. When a reason is given to explain why something happens, a certain amount of understanding is presupposed and a certain amount is provided by the reason. One might say that the framework for understanding why the phenomenon occurs is possessed, but the final piece that enables one to fit the phenomenon into that framework is missing. That piece is the knowledge provided directly by the reason.

The state of knowledge or understanding that I am trying to analyse is the state after a reason-giving explanation has been assimilated. This means that reasons may contribute to this state in different ways depending on what knowledge is presupposed before the explanation.

In this first section I aim to show that it is possible to completely understand why P occurs without knowing everything there is to know about what led to P occurring. I assume that understanding should not be analysed subjectively, in terms of a feeling of coherence for example. Rather, understanding requires actual knowledge of coherence; the feeling follows.

In the case of causal understanding - understanding *why* some phenomenon occurs - there are two kinds of knowledge that may be relevant. The first is knowing about what makes the phenomenon happen. The second is knowing a way of telling that the phenomenon would happen. These correspond to the material and theoretical aspects of causal explanation described in chapter 2.

It is pretty clear that theoretical knowledge consisting of a way of telling that the phenomenon should occur is insufficient for understanding why that phenomenon occurs. One does not understand why something happens in virtue of consulting an expert who can predict for one that it will happen. However, it looks as if having some antecedently applicable way of determining that the phenomenon should occur is at least necessary for understanding. Without it one may know

that some phenomenon was the result of some mechanism, yet not understand why that result rather than anything else occurs.

There is one simple way of resolving these apparently divergent requirements on understanding. It is to say that the material knowledge about what makes a phenomenon, P, occur only constitutes understanding of why P occurs if it involves knowing how the thing that makes P occur works, and how P fits into this working. Seen in this strong sense, material knowledge about what makes P occur provides a way of antecedently determining that P would occur.

It will emerge that understanding does not provide a means which could have been applied before P occurred of deducing with certainty that it would occur. However, it turns out that a kind of non-deductive epistemological determination is provided by understanding accounted for in this way.

An explicable phenomenon is not antecedently deducible. Rather it is antecedently intelligible. What makes something intelligible is not a way of deducing future descriptions of the world, but a way of *producing* future descriptions of the world. I contend that this is exactly what we require to satisfy the theoretical requirement on understanding. Having mastery of a way of producing future descriptions that represents the actual causal mechanism that determines some phenomenon is the key to understanding why that phenomenon occurs. There is no need to have knowledge from which it can be deduced that the phenomenon would occur.

There are two quite different kinds of knowledge that can both be described as knowledge of how a mechanism that determines some phenomenon works. Let me describe these respectively as superficial and underlying knowledge. Superficial knowledge is knowledge of the characteristic events of a process resulting in the phenomenon and of how that phenomenon fits into those characteristic events. Underlying knowledge is knowledge of the underlying nature of the process resulting in P and of why that underlying nature has the characteristic events it does.

As an illustration, consider the question, 'How does this pea-shelling machine work?' The question may concern the characteristic results of the mechanism, or it may concern why the mechanism has those characteristic results. So there are two quite different kinds of answer: 1. 'If the pods go in here and this button is pressed, peas come out here and they don't otherwise.' 2. 'There are three rotating blades driven on separate axles by a motor ...' The first way describes the characteristic results of the mechanism, while the second way describes the internal structure of the mechanism.

Understanding provided by superficial knowledge can be seen in the example of a piece of paper bursting into flames as a result of being struck by lightning. What does one need to know in order to understand that phenomenon? First of all, one should know that the paper's bursting into flames results from a lightning-burning-things-up process. (It is an exceptionally quick process - over in a flash - but a process abiding through time for all that.) So one should know that

the phenomenon is the result of a process with such and such characteristic events. This means that one must know what these characteristic events are, and one must know that the phenomenon results from a process so characterized.

Accompanying the knowledge of what the characteristic events are - i.e. of how the mechanism works in general - one should have knowledge of how the mechanism worked in this particular instance. In other words, one should know how the phenomenon - the bursting into flames of a piece of paper - fits into the characteristic type of event structure of the process that results in it. In this example one need only know that the piece of paper is located in the vicinity of the place where the lightning struck. But for other kinds of process, the knowledge may be a significant part of one's understanding.

Take the example of a computer programmed to output the sum of two numbers that are input. One wants to know why the number '8' was output. First of all, one should know how the mechanism that determines that result works. This is knowledge of the characteristic type of event structure of that mechanism (i.e. knowledge that after two numbers are input, their sum is output). In this example it is clear that one still does not know why the number '8' is output until one knows how that result fits into the characteristic type of event structure. One has this knowledge when one knows that the numbers '3' and '5' were input.

Knowledge of how a mechanism that determines some phenomenon works in general and did work in particular to determine that

phenomenon constitutes what I am calling superficial knowledge of why that phenomenon occurred. Such knowledge need not depend on any ability to identify that process reliably through its underlying and abiding nature. I might have complete superficial knowledge of why the number '8' was output if I have simply been told that the computer is outputting the sum of its inputs which were '3' and '5'. To be able to identify that process through its essential nature requires knowledge of computer science. Such ability is part of underlying knowledge of the mechanism.

My claim is that complete understanding of why a phenomenon, P, occurs is provided by superficial knowledge. I do not seek to minimize the scientific importance of underlying knowledge. What I claim is that it does not provide one with understanding of why P occurs, but rather with understanding of the process or mechanism resulting in P. I claim that reasons explaining P need not include underlying considerations. These considerations are important only in explaining the process and not in explaining P itself.

In prima facie support for this claim is the fact that there seems to be a clear difference between what it is to understand why something happens and what it is to understand the process of making it happen. One seeks reasons for why a phenomenon occurs in order to be able to fit that phenomenon into one's world picture: that is the picture built from one's knowledge of entities in the world including processes. Confusion reigns when one cannot place a phenomenon in this world of processes.

Underlying knowledge enables one to answer the next deepest 'Why?' question: why does that process work that way? Given this knowledge, one can not only place the phenomenon in the world of processes, but now also place the working of the process in the world of sub-processes. And there are always further levels of understanding to be had.

It should be pointed out that underlying knowledge does not automatically provide one with superficial knowledge. I may know all about the workings of a computer and its program, but still not understand why a certain number is output if I do not know what was input. So, having deeper knowledge does not mean that one can do without the superficial knowledge if one wants to know why P occurs.

If superficial knowledge were insufficient for understanding why P occurs, one would presumably need answers to every level of 'Why?' question in order to understand it. There is no other sensible place to draw the line. If this is as absurd a conclusion as it seems then superficial knowledge is sufficient for this kind of understanding. However, there are one or two arguments against this apparently obvious account of understanding.

Obviously, if there were no such thing as processes, then this superficial level of knowledge would not exist either. One would have to rely on the theoretical aspects of explanation. In this case, complete understanding could only be had when the theoretical way of telling what should happen was perfect. This would not be exhausted by the knowledge that I have characterized as superficial. So, this

account of understanding depends on some account of processes like the one in the previous chapter being possible.

Similarly, suppose that there were such a thing as processes but that they were metaphysically reducible to progressively finer-grained processes. By this I mean that there is a token identity between a process and the set of its sub-processes. Then it would only be a convenient way of speaking to say that a certain macro-level process was happening. It would be more complete to say that a certain pattern of sub-atomic processes was happening. In this case, knowledge of the characteristic events of the macro-level process could not be part of the real understanding of why the phenomenon occurs.

My response to this is that one should resist the thought that the process of a candle burning down is metaphysically reducible to a structure of sub-atomic processes, just as one should resist the thought that the candle itself is metaphysically reducible to a structure of sub-atomic particles. It is just not the case that these entities are identified in terms of these sub-entities. So it is always possible for the same process to consist of different sub-processes. As long as the underlying essential nature of a process is continuously present and its characteristic events occur, variation in the sub-processes is possible. This means that a token-identity cannot be set up between a process and its sub-processes.

Another argument goes as follows. Even without metaphysical reduction, it is still true that we can describe the world in terms of more and more basic processes. The more basic the processes that we

have in our world picture the more accurate is that picture. So explaining anything in terms of the characteristic events of macro-level processes is made scientifically redundant by underlying knowledge of how the world may be described in terms of the finer-grained processes.

However, there are two good reasons why superficial knowledge is not made scientifically redundant by underlying knowledge. The first is that macro-level phenomena can only be accounted for by reference to macro-level processes, unless the macro-level phenomena are themselves metaphysically reducible to fine-grained processes. Assuming that the phenomenon of a rabbit eating a blade of grass cannot be identified with some micro-level event-structure, then any explanation at the sub-atomic level will fail to explain *it*.

The second reason is that, even if macro-level phenomena could be reduced to fine-grained phenomena, the information required to make a fine-grained explanation may be practically inaccessible, whereas that required to achieve understanding on the macro-level is available. Even if it were the case that by knowing the position, velocity, etc, of every particle in a closed system one could explain some macro-level phenomenon occurring in that system, practically achievable explanation of that phenomenon in terms of superficial knowledge would be at no risk of being made redundant.

It may be that a main interest of science is in investigating and understanding the fundamental nature of processes rather than in just investigating how they work (i.e. what their results are). But this

does not make redundant every other kind of investigation and understanding. Being told R helps one understand why P occurs only if one knows the characteristic type of event structure of a process resulting in P and how R and P fit into that characteristic type of event structure together.

In conclusion, there is a level of understanding which is provided by knowledge of the characteristic events of a process resulting in a phenomenon and of how that phenomenon fits into these characteristic events. It is appropriate to describe this as understanding why the phenomenon occurs. In terms of just this kind of understanding, nothing is lacking if underlying knowledge of the mechanism is lacking. For, this underlying knowledge is part of understanding something else altogether - i.e. why the process has the characteristic events it does. This latter kind of understanding does not make the former more superficial kind of understanding redundant.

3.2 Knowledge of How a Mechanism Works

To understand why a phenomenon occurs one must know how the mechanism that resulted in that phenomenon works. One must also know that that mechanism resulted in that phenomenon and how the phenomenon fits into the characteristic events of that mechanism. Knowledge of how the mechanism works involves knowing what the characteristic events of the process of the mechanism working properly are. This

means that one must know what the different results of that process are in different circumstances.

There may be a suspicion that this requirement is too strong. One may understand full well why '8' was output by the computer when '3' and '5' were input, yet not be good enough at arithmetic to know what the characteristic output will be when '2639' and '401921' are input. However, I think this sort of case proves nothing. For, at least one knows that the sum of '2639' and '401921' will be the characteristic output.

A more interesting case is the following. Suppose that the computer has been programmed to add all input pairs except the pair '2639' and '401921' which it multiplies. Is it not unnecessarily tough to demand that knowledge of this exceptional case is required before one can understand all the other quite separate cases?

Such examples only reveal the strength of my account. Most computer programs designed to have such output would work with two separate modules, one for adding the inputs and another for multiplying them together. A third module would be used for passing control to one of the other two depending on whether or not the inputs were '2639' and '401921'. In a computer so programmed there would be an identifiable mechanism that characteristically resulted in the sum of its inputs. It would be possible to identify this mechanism quite separately from the rest of the system.

The event of '8' being output is the result of the adding process alone as well as simultaneously being the result of the combined adding

and multiplying process. These processes are neither independent of each other nor identical to each other. They happen at the same time, to the same things and with the same results. But they are as distinct from one another as are a plant and its stem or a house and its walls.

If one knows about one of the processes - the simple adding process - and not the other, then one understands why '8' was output in virtue of that. It is not a requirement of this sort of understanding that one knows about all the mechanisms that result in the phenomenon. Nor is it a requirement that one knows exactly in what circumstances the mechanism that one does know about is operating. One need only know that it is operating in this case, how it works when it does operate and how the phenomenon fits into this. So, not knowing about the fact that this adding mechanism does not work when '2639' and '401921' are input does not detract from one's understanding.

However, suppose that the computer program is not separable in this way. Suppose that it is impossible to identify a separate mechanism which always adds when it works properly. Then, in this case, one does not understand why '8' is output when '3' and '5' are input. For, the output is not the result of adding the inputs, but of a more complicated function that happens to give the same results as adding for all numbers except '2639' and '401921'. In this case, failure to know what should be the output when '2639' and '401921' are input shows a failure to understand what mechanism it is that determines that '8' is output when '3' and '5' are input.

So it seems reasonable after all to assume that one must know what the results are of a mechanism which determines the occurrence of the phenomenon to be understood in *all* circumstances within the proper working of that mechanism (i.e. when the mechanism is there, its operational conditions are satisfied and nothing interferes).

Understanding why a phenomenon occurs involves knowing what the characteristic results of a mechanism are. Given this, it may seem paradoxical that there might be two types of situation in each of which the computer has exactly the same outputs given the same inputs and one has the same ignorance about the internal nature of both mechanisms, yet in one of which one understands why '8' is output and in the other one does not. But it is not really paradoxical. In one situation one knows what are the characteristic results of a mechanism that determines the output, and in the other one does not. This is despite the fact that one's structure of beliefs may be identical in the two situations.

Now I must answer an apparently trivial question. What does knowledge of how a mechanism works involve? The question is in fact not at all trivial, since the apparently obvious answer turns out to be inadequate. It is important to have an account of such knowledge, because something is a reason in virtue of how it fits in with such knowledge. So the final account of reason-giving explanation depends on the answer.

The apparently obvious answer is this. Knowledge of how a mechanism works involves knowing the truth of a sentence of the following form:

> *If the mechanism that resulted in the phenomenon is working and nothing goes wrong, then in situation S_1 result R_1 follows, ... , in situation S_n result R_n follows.*

The kind of thinking that leads to this answer may be something like this. Understanding or knowing how something works involves knowing the principles or laws of its working. This is just like the thought that knowing how a game is played involves knowing the rules of that game. What could knowing the rules or laws of something be other than knowledge of the truth of an expression of those rules or laws? So it seems that knowing how a mechanism works must involve knowing the truth of a sentence like S.

Let us call this the Propositional Answer. This is to be contrasted with the Mastery Answer, according to which, knowing how a game is played involves having mastery of a way of telling what should follow from the rules of the game being followed. Similarly, knowing how a mechanism works involves having mastery of a way of telling what should follow from the mechanism operating. According to the approach of the Mastery Answer, understanding and knowledge are not to be analysed ultimately in terms of a subject's relationship with a proposition or a sentence, but in terms of a subject's mastery of certain methods.

According to the Mastery Answer, one knows how a mechanism works only if, in those situations where the mechanism works, one can express what its result should be. There should be an intellectual mechanism that mirrors in its operation the mechanism to be understood, except it results in the formation of a description of the result rather than in the result itself. Of course, this intellectual mechanism does not always work when the mechanism to be understood is working. But how it works, when it does work, mirrors exactly how the mechanism to be understood works.

It is possible to illustrate the difference between the Mastery Answer and the Propositional Answer in terms of a difference in scope. According to the Propositional Answer, understanding how a mechanism works involves being able to express in the present situation what the results of the mechanism working should be in all those situations in which it does work. According to the Mastery Answer, understanding how a mechanism works involves being able to express in each of the situations in which the mechanism works what the results should be. According to the Propositional Answer, the subject must be able to express a complicated thing: that is what the results of the mechanism should be in all possible situations. Whereas according to the Mastery Answer, the subject need not be able to express that complicated thing in the present situation so long as in all possible situations, they can express a relatively simple thing: that is what the result of the mechanism should be in that situation.

This seems to show the attractiveness of the Propositional Answer. Understanding how a mechanism works is a fact about the subject in their present situation. It is not a fact about what the subject would be like in other situations. However, despite appearances to the contrary, this is also the case according to the Mastery Answer. According to the Mastery Answer, whether or not someone understands how a mechanism works is an identifiable fact about them at that time. For, if a subject has mastery of a method, then there is some identifiable underlying state that the subject is in, given which the subject will exhibit that mastery in the appropriate conditions if nothing interferes. It need not be sufficient, according to the Mastery Answer, that a subject be able in various different situations to express what the results of a mechanism should be. This might rather be a consequence of the subject being in the appropriate underlying state of mastery.

Before considering the Mastery Answer further, let us put some pressure on the Propositional Answer. First of all, it is unnecessary to require that for a subject to understand how a mechanism works, the subject must know the truth of a single sentence expressing how the mechanism works. There might be no such sentence. There might be no finite specification of how the mechanism works that leaves absolutely no doubt about how the specification should be interpreted. And even if there is such a specification, a subject may not assent to it because of its great complexity (even though the subject may be able to assent to every basic condition of the specification taken individually).

In the industrial pursuit of Artificial Intelligence, computer scientists try to formulate logical specifications of how an expert does a certain task by painstakingly going through with the expert what they do in each type of situation. Assuming that it is possible to produce a correct specification, it should not be supposed that the expert knows that that specification is true. The expert might be surprised or sceptical that the specification could be abstracted from what they said. Yet there is no question that the expert knows how the task should be performed.

Given this, the Propositional Answer should be adapted so that the understanding subject only needs to know the truth of all the (possibly infinite number of) basic conditions of the form 'If situation S_i obtains, then R_i results'.

Now, one cannot know the truth of such claims unless one understands what is described by S_i in each case. One cannot understand this unless one has a way of recognizing those situations in which S_i holds true. Combining this with knowledge of the hypothetical claim 'If S_i obtains, then R_i results', means that one must have mastery of a method of producing a description of the results. So one cannot have understanding of the type expressed by the Propositional Answer without having understanding of the type expressed by the Mastery Answer.

An adherent of the Propositional Answer might ask how one is to learn how a mechanism works except by learning facts of the form, 'If situation S_i obtains, R_i results'. The answer is that there are other

ways of learning how a mechanism works. One may learn a model, and by working through that model form a description of what should follow. One may learn the use of a metaphor. One may learn how to apply certain general but vague principles like that of conservation or symmetry. One may even learn a kind of practical reasoning about what is the best thing to happen, and apply that. (This last is appropriate when understanding teleological mechanisms.) What one rarely does is learn a large number of facts of the form 'If S_i obtains, R_i results'.

So, we have reached the following conclusions. Understanding a phenomenon involves knowing how the mechanism that results in that phenomenon works, when it works properly. Knowing how a mechanism works involves mastery of a method of describing the world in terms of the characteristic events associated with that mechanism.

But what is it exactly that one has mastery of in this case? It is a method of drawing future descriptions of the world. For example, when one sees the numbers '3' and '5' being input to the computer, one derives the description '"8" will be the output'. One has complete mastery of a method of forming descriptions of the world in accordance with the characteristic events of a certain mechanism when in any situation where the mechanism is working and one has access to anything one needs to know about, one can derive the description of the result of that mechanism using that method.

Note that one can be liberal concerning the interpretation of the phrase 'forming a description'. All that is required is that the subject can somehow express how the world should be as a result of the

working of the mechanism. This expression could be in language, action or thought.

Let me use the phrase 'theoretical method' to mean a method of forming descriptions. I think that it is reasonable to assume that this is a meaningful notion. However, just how the notion of a theoretical method is grounded is still very much an open question. In particular, do we need to ground the notion of a theoretical method in some physical structure or can it stand firm on logical ground alone?

One approach is that of Wittgenstein who sought to ground the notion of a theoretical method in the linguistic practice of producing descriptions. Talk of such methods is abstracted from talk of such practices. An alternative approach, rejected by Wittgenstein, is to ground the notion of a theoretical method in that of a *process* of producing descriptions. Let me elaborate briefly on this second way of grounding the notion of a theoretical method in a physical structure.

A farmer looks at the sky and says that it will rain overnight. Their description is the result of some intellectual process; and underlying this process is some mechanism. To go from this to the abstract notion of a theoretical method we need to introduce the possibility of a description being *theoretically determined* independently of whether it is causally determined by such a mechanism. This is to be done in terms of *potential* causal determination. So, we say that a description is theoretically determined if it *would* be determined by the intellectual mechanism. Out of this we can abstract

the notion of a theoretical method. The theoretical method is that which theoretically determines descriptions.

The state of understanding how a mechanism works can be now accounted for. The formation of a description R_i is causally determined by another mechanism; that is the intellectual mechanism of forming descriptions of the results of the original mechanism. This intellectual mechanism may be described in terms of a method. We say that in situation S_i the description R_i is theoretically determined by the method, when the formation of such a description would be causally determined by such a mechanism if it were operating.

The state of understanding is the state of having nearly all the preconditions for the intellectual mechanism to be working. What is not required is that the subject is in a position in which the information concerning the present situation is available. Also there is no need for the subject to have good reason to produce a description of the results of the mechanism. These are part of the underlying nature of a description forming process, but are quite beside the point as far as understanding how the mechanism works is concerned. If one is in this state, then if one is given access to the present condition of the world and one has a reason to form a description of the future world, then one produces the correct description R_i.

So, understanding how a mechanism works involves having the makings of another mechanism (an intellectual mechanism) that would determine descriptions being formed of the results of the original mechanism. If one had understanding in this sense, but could not

express how one produced the descriptions, then the Mastery Answer would be satisfied but the Propositional Answer would not be satisfied. I think that we would have no problem in still describing such a state as that of understanding.

Now, a theoretical method, T, is supposed to describe how a causal mechanism, C, works. Just how does it do this? In different types of circumstances C itself will have different results. We say that T represents C if and only if in all the different circumstances, where C is working properly, its results are those described by the results of T (also assumed to be working properly). In other words, how the mechanism works is reflected by how the method works.

For example, consider the mechanism of Newtonian gravitation. (I am assuming there is such a mechanism.) Newton's theory of gravity provides descriptions of how things will interact which truly describe the results of the gravitational mechanism. Whenever the Newtonian gravitation mechanism is working its results are what are described by the Newtonian theory of gravity.

This is not to say that in all circumstances the theory must determine descriptions that are true. For, there is no need to consider those circumstances where the mechanism is not working properly. This is a consequence of the earlier result that superficial knowledge of how a mechanism works is all that is required for explaining a phenomenon. One does not need to have the underlying knowledge of the exact circumstances in which the mechanism is operating.

So, for a theoretical method to be explanatory it does not have to involve universally true laws. It must result in descriptions that turn out to be true - *but only when the mechanism that is being described is working properly and nothing goes wrong*. And this is a much weaker claim.

In strange Einsteinian situations Newton's law of gravity breaks down. These may be regarded as situations in which the Newtoninan mechanism of gravitation does not work (and must be included in our notion of what the Newtonian mechanism is). This does not mean that Newton's theory is wrong. The worst we can say about it is that it does not embody universally true laws.

3.3 Intelligibilty and Theoretical Predetermination

By intelligibilty I mean what is left of understanding when causal determination is taken away. On the Humean account this will be the whole package; on mine it should turn out to be something much less. I consider what notion of intelligibilty is provided by my account and whether it is adequate.

Explaining something makes it intelligible. However, there is a sense in which something may be intelligible without being explained. In this sense whatever makes something intelligible can occur without the underlying process of causal determination. For example, one can make the ignition of a piece of paper intelligible by telling the story

of the lit match placed under it, even when the true explanation of the ignition of the piece of paper has got nothing to do with the lit match. I want to capture this epistemologically weak notion of intelligibility.

On the present account we have the simple conclusion that a phenomenon is intelligible if its description is theoretically determined by a method that is mastered by the subject and which might have represented an actual mechanism. That one has mastered a theoretical method does not mean that one has any knowledge that could have antecedently enabled one to tell that the phenomenon would happen. Nor even need one be in a position in which it would have been reasonable to assert that the phenomenon would happen.

So we have a very weak notion of intelligibilty. Although a theoretical method that underlies theoretical determination is available antecedently and the description that the phenomenon will occur is determined by it, yet there is no *implicit* guarantee that the method is the right method. We are very far from the 'covering-law' idea that whatever makes something intelligible provides antecedently a guarantee that it will come about.

A future tense sentence may be determined by one method while it is contradicted by another. What it is reasonable to assert is determined by a wider consideration than just what description should be made according to one method. If a way of determining future descriptions determines the description that X will occur, we cannot thereby say that it is reasonable to assert that X will occur.

However, we can say that a *conditional* assertion is rationally justified; i.e. 'According to this way of drawing descriptions, it *should* be the case that X occurs.' This is a kind of epistemological predetermination however weak.

This fits in with the notion of understanding based on superficial knowledge fixed in 3.1. Such understanding does not in itself mean that one has antecedently applicable reasons for believing that the phenomenon would occur. If knowledge of interfering conditions is not necessary for such understanding, then one need not be in a knowledge position that would give one confidence that the phenomenon would occur antecedently. Certainly one does not have reason from which one could have *deduced* that the phenomenon would occur.

Moreover understanding why a phenomenon happens does not mean that one has the ability to identify the process antecedently. This means that one lacks even antecedent reason to believe that the phenomenon will happen if nothing goes wrong. For one's knowledge that a phenomenon is the result of a certain sort of process which underlies this sort of understanding is not something that one need have before the phenomenon occurs.

So being equipped with understanding of why a phenomenon occurs does not mean that one is equipped with anything that would have enabled one to conclude rationally that the phenomenon would occur in advance. What one is equipped with is something that enables one to determine in advance the *idea* that the phenomenon will occur. This

only enables one to conclude that the phenomenon should occur according to that way of describing things.

There may be worries now that such an epistemologically weak notion of intelligibilty allows almost anything to be intelligible. If theoretical determination is not logical deduction, what is to stop it from being arbitrary and ridiculous? Must we impose constraints on what methods of forming descriptions may count as making things intelligible?

My answer is no! To begin with, nothing can count as a method of describing the world if its results are not consistent. So there is an inbuilt constraint of consistency in the very idea of a theoretical method. 'P and not-P' can never be part of a description of anything. (So it is with games. A set of rules would not identify a game if they determined contradictory instructions as to how to act.) So not just any set of sentences can be produced by a theoretical method.

A second source of inbuilt constraints would be available if theoretical methods did require grounding in intellectual processes as was suggested in 3.1. For then, something could not be a theoretical method and be totally arbitrary. Randomly making up descriptions as one goes along is not a method that can express a *determinate process*.

Does this not still leave us with the possibility that some very crazy world picture would be intelligible? Yes, and this is the way it should be; for it is not the task of theoretical determination to make a necessarily right description. It is in the relation between the theoretical method and the real causal determination that the

correctness or incorrectness of a way of describing the world is decided.

Consider a crazy theoretical method that incorporates the rule 'Whenever someone walks under a ladder determine the description that ill luck will befall them'. This is at least a method of forming descriptions (as is indicated by the fact that people actually use it). It should not be discarded as a way of making something intelligible. (Remember the epistemologically weak sense of the word 'intelligible' that makes it closer to 'assimilable' than 'understandable'.) For if there was a mechanism represented by this sort of superstitious method then the method could be used to explain cases of ill luck.

So it is the world that does the work and not the theoretical method in distinguishing crazy explanations from good ones.

One last possibility needs to be looked into. Might there be processes that involved indeterminacy - processes that allow different results in the same circumstances? These would be represented by theoretical methods that could determine descriptions of any of the possible outcomes.

A simple example would be the process of rolling a die. The result of this process might be that a 6 is shown uppermost. But, there are five other possible results given the same circumstances (the same as far as the characteristic events of the ~~the~~ process can discriminate). A theoretical method that represents this would be to pick a number between 1 and 6. This method would be indifferent between the six possible results. The process is also indifferent between these

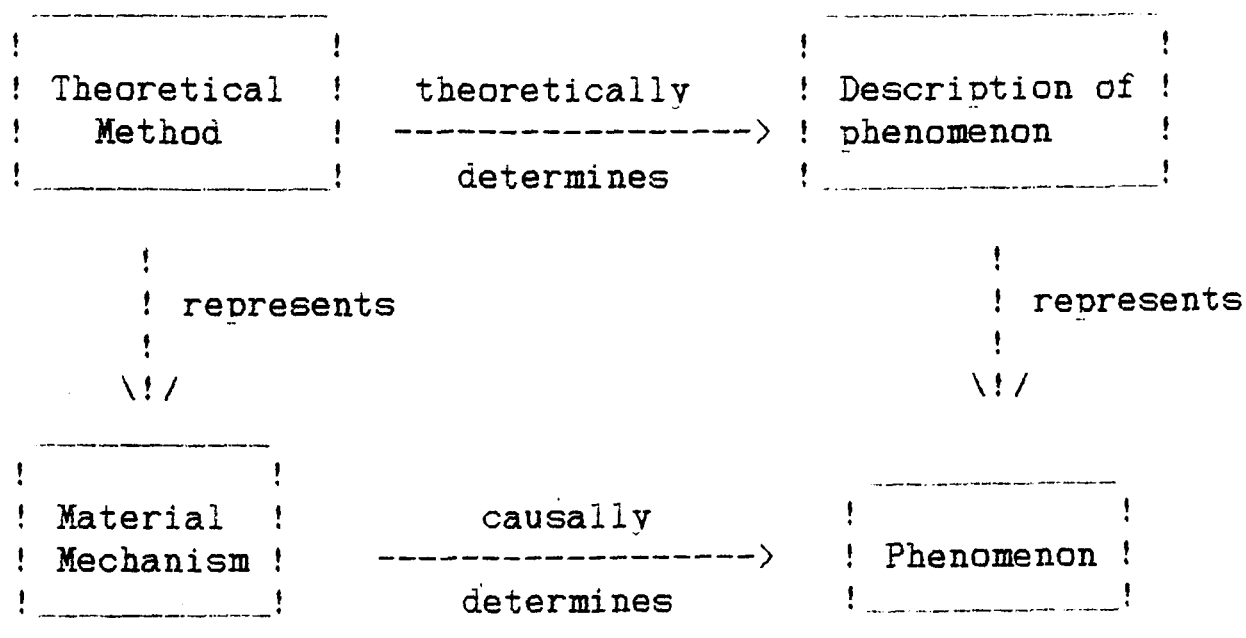
results. So, it looks as if the theoretical method does represent the process.

If one did not want to allow the possibility of processes involving indeterminacy, this example could be treated differently. One might say that a 6 showing uppermost is not the result of that process. Instead, the process results in a number between 1 and 6 showing uppermost. So, instead of the process having an indeterminate result, it just has a less precise result. At the same time, there is another process which does result in a 6 showing uppermost, but this process picks up the smallest details of the situation and can only be represented by a much more complex theoretical method.

But now consider the process of getting a total from the rolls of two dice. It is no longer satisfactory to specify imprecisely a determinate result of such a process. To say that the process results in a number between 2 and 12 being totalled misses something about the result; which is that 7 is more likely than 2. The process is not indifferent between these two results. The theoretical method consisting of picking a number between 1 and 6 twice and adding them is not indifferent between the results of 2 and 7. So, it seems to represent the process more completely.

3.4 Causal Explanation

I have now completed the derivation of the Square of Explanation.



In order to understand why phenomenon, P, occurs one must have the theoretical knowledge shown in the top half of the Square, and this must represent the actual process of causal determination resulting in P, shown in the bottom half. One must know how the mechanism resulting in P works - i.e. one must have the theoretical method. And one must know how the theoretical method determines that P should occur.

It is worth mentioning in passing that *knowledge* as such is not required for understanding. Suppose that one believes that some theoretical method represents a mechanism resulting in P. And suppose that this belief is true only by accident. Then one cannot be said to *know* that that method represents a mechanism resulting in P. Yet one

can still be said to understand why P occurs. The mental state of understanding is that of a capacity to fit P in correctly with the world of processes. There is no need for the capacity itself to be reliably arrived at.

The next task is to analyse the notion of reason-giving explanation in terms of this model of what it is to understand why P occurs.

The simplest account is that an explanation of P is that knowledge (or belief) one must have in order to understand why P occurs. But this ignores the fact that an explanation is essentially something that is provided in response to a plea for understanding. An explanation is a linguistic act.

Achinstein (1983) develops an illocutionary theory of explanation. According to this, an explanation is something you do when you act with the intention of making somebody understand. An alternative account might regard explanation as a performance act - i.e. it is something you do when you succeed in making somebody understand.

Both of these linguistic act accounts suffer from the criticism that explanatorily irrelevant things may be said as part of an attempt to make someone understand. One may have to revert to the simple account to determine what is explanatorily relevant. This would mean that the fact that explanation is a linguistic act is not analytically important.

Given any of these accounts of explanation we can show what a reason is. Given the simplest account of explanation, a reason why P occurs is some fact you have to know in order to understand why P

occurs. Given the linguistic act account, a reason is a fact provided in an explanatorily relevant attempt to make someone understand why P occurs.

Exactly the same things will come out as reasons in both types of account. The only difference is that things which are reasons according to the simple account may only be potential reasons according to the linguistic act account. The condition of explanatory relevance means that all reasons by a linguistic act account must be reasons by the simple account. Conversely, if some fact must be known if one is to understand why P occurs, then the only way to make someone who does not know this fact understand why P occurs is to tell them it. So that fact is potentially a reason by the linguistic act account also.

Given all this, it is clear that there are two quite distinct types of reason that may be provided in an explanation of why P occurs. The first sort of reason concerns the way a theoretical method representing a mechanism resulting in P works. According to the account I provided in 3.2, the appropriate sort of knowledge here is practical knowledge. This sort of knowledge does not consist in knowing a set of reasons. So, although reasons may usefully be provided here, they will not completely constitute this part of an explanation.

The other sort of reason concerns the way P fits into the characteristic events of the process which results in it. How the process resulting in P works is taken for granted. For example, in this sense, a reason why '8' was output is that '5' and '3' were input.

Whereas, in the first sense, one reason that '8' was output is that an adding process was happening.

I am only concerned with reasons in this second sense. We can think of explanation as having two parts. The first part consists of teaching the correct theoretical method. The second part consists of providing a set of facts that enables the learner to see how the theoretical method determines that P should occur. I will only use the word 'reason' to refer to facts required for this second stage.

It is a consequence of this usage that one may know all the reasons why P occurred without understanding why P occurred, if one does not know the correct theoretical method. This usage does not seem too unnatural. It does seem natural for us to say that one reason that '8' was output was that an adding process was happening. But this may be because there is a meta-process operating here, where the fact that an adding process was going on is one of the reasons in my sense. This meta-process has different results depending on what program is input.

In any case, it does not really matter whether my use of the word 'reason' corresponds to common usage. I am quite happy for it to be regarded as a technical term for present purposes.

Given this, the formal requirement for a description R to be a reason that a phenomenon P occurs is that a theoretical method that supplies a correct explanation of why P occurs would not intellectually determine a description of P if R were false.

For example, the fact that lightning struck the piece of paper is a reason that the paper ignited. Had the lightning not struck, we would

not be able to apply our theoretical method to provide the description that the paper ignited.

Suppose that it is the case that had the lightning not struck, the piece of paper would have ignited anyway, since I was just about to light it with a match. We might still say in this case that the fact that the lightning hit the piece of paper is the reason it ignited. This is because there is a theoretical method that just represents how the lightning-as-burner mechanism works, and according to this method, the lightning must strike the piece of paper if we are to determine that the piece of paper ignites.

Now suppose instead that if that bolt of lightning had not struck the piece of paper another one would have done. It is not the case that the fact that that particular bolt of lightning hit the piece of paper is the reason why the piece of paper ignited, even though we can say that it is the reason why the piece of paper ignited just when it did. For we have no method of determining the description that the piece of paper ignites which depends on that particular fact.

We can now formulate an analysis of explanation.

Phenomenon P occurs because of reason R if and only if there is a theoretical method, T, that determines a description of P and represents the characteristic events of a process that determines P itself; and T would not determine a description of P if R were false.

Consider the kinetic theory of gasses. Application of the theory gives descriptions of temperature, pressure, etc of a gas. When the kinetic mechanism is working properly, the kinetic theory gives true

descriptions of the results of the mechanism. Thus the kinetic theory represents the kinetic mechanism.

It may be argued that the kinetic theory is in fact a bad (incorrect) theoretical method. It treats molecules as little billiard balls bouncing off each other: and this is wrong. But this argument, far from damaging our theory of explanation reveals the strength of it. In as much as the kinetic theory determines descriptions of the nature of molecules as hard balls it does not represent what is actually happening. But when the kinetic theory is used more schematically with a vagueness about the physical nature of the molecules and their interaction ('they interact *as if* they were billiard balls') then the theory does not determine false descriptions. When the theory is used this way, with no literal implications about the nature of molecules, it is regarded by scientists as a good explanation.

Another sceptical argument that can be brought to bear on this example asserts that there is in fact no kinetic mechanism. There is a kinetic theory that sometimes gives true descriptions as if there were an underlying kinetic mechanism. But considered as a totally general fact, it is not the case that the kinetic theory gives true descriptions. At very high pressure for example its resulting descriptions are all false. It seems that whatever underlying mechanism is at work must be something more subtle.

This argument depends on the thought that there is a unique mechanism underlying any phenomenon. The correct explanation must then be the one that describes this mechanism. In opposition to this,

I have argued that there may be different and mutually incompatible explanation structures representing quite different mechanisms, any of which can be seen (taken separately) to be determining the phenomenon. Some mechanism other than the kinetic one determines the behaviour of a gas at very high pressures. Presumably, this also determines the behaviour at normal pressures. But this does not preclude the kinetic mechanism from also determining the behaviour of the gas at normal pressures.

This does not mean that I allow that a theory may determine a mechanism. And it does not mean that, according to my account, a mechanism is really a theoretical structure rather than something out there in the physical world. What is determined by theory or by the way we use language is what part of the physical world is to be regarded as the proper working of a particular mechanism (its characteristic structure of events).

Consider the following attempted explanation. At certain magic positions of the sun and moon the great sea spirit decides to deposit seaweed on the pavements of the seaside town, rises up into a great flood tide, makes the sea drop seaweed on the pavements and then recede satisfied. Is my account of explanation so liberal as to allow this?

The first problem with this attempt at an explanation is that it yields false descriptions even when the mechanism it should represent is working. For example, it determines the description that there is a sentient being in control of the sea. So, let us change the explanation

to 'It is as if there is a great sea spirit ... '. Given this change, the explanation is now a florid way of saying that at certain positions of the sun and moon the sea rises in a great flood tide and seaweed is dropped on the pavements as the sea recedes.

At first sight we might say that this theoretical structure represents a simple flood tide mechanism of the sea at this seaside town. It represents the very simple mechanism of the sun, moon and sea that works in the following way: when sun and moon are in position X, the sea rises up and drops seaweed on the pavements; otherwise it doesn't.

However there is a potentially worrying aspect to this example. For what has been described is not in fact a flood tide mechanism but a mechanism for dropping seaweed on the pavements. According to my account of explanation, I cannot back out of this conclusion by arguing that no such mechanism is at work for if the town built a sea wall the sea would still rise but not drop seaweed on the pavements. For the correct response is that if a sea wall is built then the mechanism for dropping seaweed on the pavements does not work properly; but as it is, without the sea wall it does work properly.

Instead my response is to embrace the apparently unwelcome conclusion. A mechanism for dropping seaweed on the pavements is at work. If we want to talk like this we can. Why we do not usually want to talk in this way is that the theoretical method that represents such a mechanism has such a small applicability.

The notion of dependence may be defined quite simply in terms of determination. Consider this definition.

How things are concerning aspect A depends on how things are concerning aspect B if and only if how things are concerning aspect A is determined by a process, the characteristic events of which can be described as a function of how things are concerning aspect B.

For example:

1. The way billiard ball B moved depended on the way billiard ball A was struck.

2. The colour of the newspaper depends on its age.

3. Your income depends on the number of years service.

My account of explanation is this:

How things are concerning A is causally explained according to a theoretical method, T, if and only if how things are concerning A is determined by a mechanism that is represented by that theoretical method.

A mechanism is represented by a theoretical method if and only if the results of that mechanism working properly are described by the results of the theoretical method being applied properly. If we put this together with the definition of 'dependence' we get:

How things are concerning A is explained according to a theoretical method if and only if how things are concerning A depends on what is intelligible according to that method.

So we might say that what happens to a bit of paper struck by lightning depends on what is intelligible according to a scientific method involving theories about lightning. It sounds from this as if

it follows that what happens to the bit of paper depends on how we apply the scientific method. This of course is an undesirable consequence implying that human intellectual activity has somehow got involved in the causal lightning proces.

Fortunately this undesirable consequence does not follow. The intelligibility that the activity of the piece of paper depends on is independent of this or that person finding it intelligible. This is a confusion that it is particularly important to avoid when the account is used for teleological explanation. The plausibility of saying of teleologically explicable activity that it depends on what is intelligible according to practical rationality must not lead us to conclude that teleologically explicable activity must be determined by a process of intellectual rationalization.

3.5 Applications

The purpose of science is to improve theoretical methods; to make them more comprehensive. The inadequacy of science at any one time is that it fails to account for everything, not that it accounts for everything wrongly. The latter idea is associated with identifying science with a body of facts (including laws) most of which are slightly false. Improving explanations is supposed to involve getting nearer the truth. But there is no such thing as a slightly false fact.

and there is no measure of 'nearness' to truth; if a description is not true it is false.

A set of scientific facts answers questions and generates descriptions. It is an active structure. Its explanatory success is not to be measured by its truth quotient (which must be 100% or it is not a set of facts), but by the comprehensiveness of the set of questions answered by the structure.

Consider causal explanation outside the physical sciences. First let us look at historical explanation. Consider an explanation of why the Church of England was established as the key religious institution in England. This will consist in a story describing Henry VIII's battle with Rome over his divorce, the dissolution of the monasteries, etc. We want to say that if the story is well told there is something inevitable about the last part of it - about the establishment of the Church of England. This inevitability is certainly not logical necessity. It is the inevitability of the result of the correct application of a theoretical method.

If one grasps an historical explanation one must be able to use it to determine other descriptions than the one at issue, even if only the contradictory of the one at issue. 'If the pope had not been in league with Philip of Spain, then he would have been happy to anull Henry's marriage, and the Catholic church would have remained in domination in England for some time.' Again note that we are not deriving conclusions from universally true laws. There is no need to rule out

the logical possibility of some other cause for confrontation with Rome cropping up unexpectedly.

What the historical narrative does is to identify and describe a process that resulted in the establishment of the Church of England. And it identifies the process by describing how the underlying mechanism worked. Implicit in the historical narrative is a theoretical method that determines descriptions of what should happen in different circumstances. In this situation it determines the description that the Church of England is established. This is what gives the conclusion its feeling of inevitability. In other circumstances other descriptions would be determined. The narrative is not just a series of facts about what happened; it also embodies a series of counterfactual conditionals that enable one to produce descriptions of what would have happened in other circumstances.

The narrative describes a mechanism, a causally active structure consisting of the relevant characters, institutions, etc which works in the way described in the historical narrative. If the mechanism were put into action in different circumstances, the results would be those described by the results of the theoretical method implicit in the historical narrative in those circumstances. The method need not work for the complete range of circumstances, but only for those where the mechanism is regarded as working properly.

The final point about the historical explanation is that the mechanism described must in fact have resulted in the establishment of the Church of England. If new documents emerge that show that the

Church of England had in fact been established well before Henry VIII, then the historical story we have been considering, however compelling, fails to explain the establishment of the Church of England. There must be a theoretical aspect and a material aspect, and the theoretical aspect must represent the material aspect (the causal mechanism) in its own working.

The account works for all types of explanations. Consider an economic explanation. Suppose one explains the increase in exchange rate with an economic model. The application of this model results in descriptions about the exchange rate. It is a theoretical method. Of course, it is not itself what determines how things are concerning the exchange rate. But it may represent the way a mechanism works that does determine the increase in the exchange rate.

What I am most interested in in this thesis is teleological explanation. I shall argue that some piece of activity may be made intelligible by a theoretical method that concerns the desirability of the likely results of that activity. If that theoretical method represents a mechanism that in fact causally determines the activity, then the activity is explained according to that teleological method. It will turn out to be crucial to my account to maintain the clear separation argued for in these chapters on explanation between the teleological method of theoretical determination which results in descriptions about future activity and the causal mechanism that is being represented by that method.

Chapter 4

Teleological Explanation

4.1 Introduction

The key principle of the entire thesis is this:

Key Principle: *If something can be causally explained in terms of its being objectively practically rational, then it may be said to be purposeful.*

For example, if P happens because it is the best way of achieving a result, G, then P happens with the purpose of achieving G. 'Purposeful' is to be understood in a strong sense here. It is not to be equated with 'having a function'. According to this strong sense, once we have an account of purposefulness, nothing else needs to be introduced into the story before we can see on what basis beliefs, intentions, etc are ascribed to people.

The Aristotelian conception of teleological explanation introduced in 1.4 was defined by the following three claims.

Claim 1. Teleological explanation is essentially explanation in terms of a practical justification.

Claim 2. There is a central notion of teleological explanation that is externalist.

Claim 3. Teleological explanation is a kind of causal explanation.

Given this conception, the Key Principle is this.

If something can be teleologically explained then it is purposeful.

Leaving open for the moment the question of whether anything ever could be teleologically explained in this sense, let me flesh out the notion of teleological explanation in terms of the account of causal explanation provided in the previous chapter.

Something is to be teleologically explained when it is causally determined by a mechanism the proper working of which is represented by a theoretical method embodying objective practical rationality. I.e. something is teleologically explained when it depends on what is objectively practically rational.

Given this, the Key Principle is that something that depends on what is objectively practically rational is thereby purposeful. This is a satisfyingly simple and straightforward theory. The idea is that if you make some activity intelligible by placing it in an objective structure of practical reasons and if the activity causally depends on what is intelligible according to that structure, then it is purposeful. Some phenomena depend on what is intelligible according to some sort of theoretical rationality. Some may depend on what is intelligible according to some sort of practical rationality. These latter are the teleological phenomena. (As will become clear later, there is no objection to the same phenomena being explained in both ways.)

How exactly to flesh out this notion of practical rationality is the subject of chapter 5. But the rough idea is this. A theoretical method that embodies practical rationality works by means-ends

analysis of some sort. Imagine a method of describing what is best or most desirable. This method will have to go from describing something as best to describing as best the best way of achieving that thing. Then find a mechanism that is represented by that method. That mechanism responds to varying circumstances (assuming it is working properly) by adapting its results to whatever is described as best in the circumstances according to that method. Now, according to the Key Principle, you have a purposeful system.

There are two stages in the derivation of the Key Principle. In the first stage I need to show that the Aristotelian conception of teleological explanation satisfying the three claims is a real notion of causal explanation that is appropriate for explaining human purposeful activity. In the second stage I need to show that anything that can be so explained is purposeful in the strong sense.

In this chapter I aim to achieve half of the first stage. I argue that the Aristotelian claims 1 and 3 can be put together and are both in play in explanations of human behaviour. The reason for limiting myself to these two claims at this stage is to keep internalists happy until the last possible moment. I see no reason why someone who accepts the internalist shift for the explanation of action need object to my account of teleological explanation at this stage. They need only drop the requirement that the practical rationality that teleological explanation depends on must be objective.

An internalist may accept that something is teleologically explained when it causally depends on what is intelligible according to

Humean practical rationality. They may accept that this sort of explanation is appropriate for human behaviour. They should also accept the second stage in my argument, that anything that is teleologically explained is thereby purposeful. For an internalist this will be the relatively trivial thesis that anything that is explained in terms of beliefs, desires, intentions, etc according to Humean practical rationality is purposeful. (As I show later, I have given the internalists a way of avoiding the problem of deviant causal chains.)

So, in this chapter, I am not insisting that the Aristotelian claim 2 must be part of teleological explanation. Nevertheless, when I argue for the second stage - that anything teleologically explainable is purposeful - I will be assuming claim 2 might be part of the notion of teleological explanation. So, in chapter 5, when I complete my argument for the first stage by introducing claim 2 into the account, I will have derived the complete externalist version of the Key Principle.

The sort of objectors I envisage to my first stage as presented in this chapter are people like Stich (1983) who argue that folk psychology is a bad causal explanatory structure. They argue that there is no such thing as *causal* teleological explanation. This is supposed to have been shown by scientific advance since Aristotle's day. It may be granted that there is some pragmatic benefit in using a teleological format in explaining things, but the real explanatory work can have nothing to do with teleology.

According to this objection, there may be some sense in talking teleologically about phenomena. This sort of talk works because it is

as if the phenomena should be explained in terms of practical rationality. However, according to this objection, there can be no real teleological mechanisms; practical rationality, whether subjective or objective, cannot actually get involved in the process of causal determination. I answer this sort of objection in 4.2.

The second sort of objection is with the second stage of my argument. This objection may allow that there is room for real causal teleological explanation satisfying the three Aristotelian claims, but argue that it is not a strong enough condition to account for purposefulness. According to this objection, teleological explanation may have a place in nature - in animal behaviour and evolution for example - but that there is a vast gulf between such natural phenomena and human purposefulness. According to this objection, the notion of teleology has exhausted its power to account for things once it has accounted for natural functions; without bringing something like mental content into the account such a notion cannot by itself account for real purpose.

More extreme versions of this objection may claim that the behaviour of smart machines or even very stupid systems like U-shaped tubes filled with water are also teleologically explainable. If any of these claims are true, then we cannot account for purposefulness in terms of this notion of teleological explanation.

To neutralize the persuasive force of this objection, I must account for the notions of natural and artificial function and show that they are weaker than the notion of teleological explanation. So, in 4.5 I

analyse the nature of evolutionary explanations and show how they fall short of proper teleological explanations. And I do this without introducing any elements like purpose or intention into the account of teleology. In 4.4 I show why the behaviour of smart machines is not teleological despite appearances to the contrary. And in 4.3 I show why simple equilibrium-directed acitivity, like that of water in a U-shaped tube, does not count as teleological, given the Aristotelian conception.

There is a powerful intuition undermining my account, which is that human intentionality cannot be accounted for as easily as that. Just to avoid knee-jerk reaction against it, I should make it clear what sort of account this is *not*. It is neither a type- nor a token-physicalist account. I am not claiming that it is possible to reduce any claim about a person's mental states to a claim about the mechanisms specified in physicalist terms that determine a person's behaviour. Instead, I am claiming that it is possible to reduce any claim about a person's mental states to a claim about the mechanisms determining their behaviour. But these mechanisms can only be specified as practical rationality mechanisms of some sort. And such talk cannot be reduced to physicalist terminology. Nor will there be any physicalist description of how such mechanisms work.

The point is that this requirement of teleological explainability is a very strong requirement. That is why nothing can satisfy it without being purposeful. A practical rationality mechanism is a mechanism that is represented by a theoretical method that embodies practical

rationality. Just what this involves is clarified later. But it is easy to show just how powerful such a mechanism is. Suppose that a practical rationality mechanism, Mechanism-1, is determining my behaviour. Suppose that the theoretical method that represents how Mechanism-1 works determines the description that G will occur, and thereby determines the description that A will happen since A is the best way of making G occur. Suppose for the sake of argument that G is Mr X getting a letter from me, and A is the process of my posting a letter to Mr X. Given that A has happened, another mechanism, Mechanism-2, determines G. Mechanism-2 is the mechanism of postal delivery. It is not a teleological mechanism.

What is interesting about Mechanism-1, the practical rationality mechanism that determines A, my posting the letter, is that it also determines G, Mr X getting the letter. For, the theoretical method that represents Mechanism-1 determines the description that Mr X will get a letter; so if the mechanism is working, it will result in Mr X getting the letter. But does this mean that the teleological mechanism must somehow incorporate the postal mechanism? This would be a disasterous conclusion. But, if Mechanism-1 does not incorporate Mechanism-2, how can it determine G which we know is determined by Mechanism-2?

For simplicity, suppose that we can describe how Mechanism-2 works with the formula 'In circumstance C_i, G_i occurs', where the C_i and G_i are appropriately specified. So Mechanism-2 gives different results depending on which of the C_i holds. This is not true for Mechanism-1. Independently of which C_i holds, Mechanism-1 will determine G. This

seems paradoxical, but it isn't. For, whether Mechanism-2 operates at all rather than some other mechanism depends on which of the C_i holds. Mechanism-1 *is* sensitive to which C_i holds in its determination of A. In a circumstance in which Mechanism-2 would not determine G, Mechanism-1 determines a different intermediate event A' so that Mechanism-2 is no longer operating. Some other mechanism that will determine G in the circumstances is brought into operation.

If C_i is that Mr X does not have a postal address, then Mechanism-1 will not determine my posting the letter, but instead perhaps will determine my giving someone the letter to take round to Mr X. My mechanism for achieving G does not have to embody the postal service, instead it must embody sensitivity to which mechanism is best to achieve G. Given this sensitivity, G will naturally follow unless anything interferes.

My claim now is that a system whose activity is determined by a practical rationality mechanism has such a powerful capacity that it is thereby capable of perception and reasoning. When the system adapts to new sets of circumstances with activity that is explained with respect to new versions of what is best given these new circumstances, we can describe the system as working out or reasoning towards the best activity. We can also describe it as perceiving what is the best thing to do and of perceiving in practical terms the details of the new circumstances. For, the system has a way of adapting to what is the most *likely* way to achieve some result; and this is a conceptual

capacity. Such a system has goals. Its activity has purpose. It is an agent.

Evidently this claim is very difficult to prove. But a story might make it more convincing. Imagine a new life form or machine that appears suddenly in your back garden - the Blob. It has moving parts. You begin to find patterns in the movements of these parts. Because of a complex pattern of movements of the pseudopodia the Blob travels along the grass. You pick it up and turn it upside down. The pattern adapts, and the Blob continues to move along. You drop it in the swimming pool. It sinks to the bottom; then after a while a new pattern of movements emerges and it flaps its way through the water, seems to pull itself over the edge and continue travelling.

So far the Blob has shown the adaptability of a simple animal. Purpose cannot sensibly be ascribed yet. Soon though, you think you see a direction to its travels. There is a bowl of milk that you have left out for the cat. If you move the bowl, the Blob's direction also alters. You let the Blob get very close and then you pick up the bowl and move it elsewhere. After you do this a couple of times the Blob's behaviour pattern alters. It travels in the opposite direction to a patch of bamboo where, because its pseudopodia are twisted a certain way around a stem, that stem breaks off.

Now the Blob has a six foot long stem of bamboo sticking out of it. It sets off again, but in a more roundabout route which seems to have no direction. You want to see what it is up to. Sometimes it seems to be going towards the bowl of milk, but then it veers off. On

one such occasion it does not veer off and gets within four or five feet of the bowl of milk. You quickly go over to move the bowl. But you cannot get close to it, for you receive an almighty jab in the shins from the bamboo spear. The Blob continues to move closer to the bowl, but now if you get within six feet of the Blob you get the sharp end of the bamboo spear poked into you. The Blob eventually gets to the bowl and ingests the milk. You stare into each other's eyes thoughtfully.

The purpose of this story is to show that we ascribe agency to something on the basis of the adaptability of that thing's activity to what is practically rational in the circumstances. You do not know exactly how the mechanism that is determining the Blob's activity works. But when you infer from the pattern of activity that the mechanism determines whatever activity is the best way in the circumstances of making some other state pertain, then you ascribe purpose and agency. If there is real strategy, then there is purpose.

Note, I am not suggesting that it is right to ascribe purpose to the Blob just because it feels natural to do so. There is a clear distinction between a practical rationality mechanism and a mechanism that results in activity that seems rational. So the rather arbitrary basis for ascribing mental states formulated by Dennett for example is not enough on my account. The system must actually be driven by rationality, not just seem to be so driven.

There are many potential difficulties with this account. Must the goals be fixed? Can a practical rationality mechanism fail and how can

one tell the difference between that and there being no real practical rationality mechanism there at all? What does it mean for something to be described as best? But all these potential difficulties will be dealt with. The main point that I need to present at this stage is that we can explain some system's activity teleologically without having to have a prior conception of that system being purposeful; but that once we can explain its activity this way, then we can ascribe purpose.

4.2 Is teleological explanation a kind of causal explanation?

In 1.1 I argued that what is demanded of a teleological explanation is a causal explanation. So, if an attempt at a teleological explanation is not a causal explanation, it fails. The question remains, however, whether such an explanation is ever possible. Perhaps all such attempts must fail. (The question of whether teleological explanation is externalist remains open through this section.) In this section I consider various arguments that purport to show this.

First of all there may be a worry that if teleological explanation is a kind of causal explanation, then this means that causation can go backwards in time. If a phenomenon is causally explained in a way that is represented by means-ends analysis, then it may be causally explained in terms of the achievment of a later goal. Does this not

mean that the achievment of the later goal causally influences the phenomenon?

The internalist response is that all that is required is that a *representation* of the goal explains the phenomenon. This response is immune to the backwards causation problem. But for an externalist, no such response is possible.

For an externalist, the consequence of backwards causation depends on treating the later goal with respect to which some activity is explained as a *particular* future state. If we allow that some activity occurs because a particular future state follows from it, and we assume this to be a causal explanation, then we must conclude that the future state causally influences the prior activity. But this can be avoided by realizing that teleological activity is explained with respect to a general type of future state rather than a particular token of a future state. So we can say that some teleological activity occurs because in general a certain type of result follows from it without in any way implicating backwards causation.

At the most general level, what distinguishes a teleological theoretical method is this. The method of forming descriptions that represents a teleological mechanism (on the assumption that there are such mechanisms) must determine the description that the teleological phenomenon will occur in a way that depends on being able to determine the description of a state (the goal state) that *comes after* the phenomenon. A teleological theoretical method must be able to arrive

at a description of the result through first having overshot the result and then come backwards to it.

For example, if one can determine the description that someone will satisfy their thirst, one may thereby determine the description that they will get themselves a drink. If the mechanism that actually results in their getting themselves a drink is represented by such a method, then the result may be explained in terms of something that comes after it.

In such cases we may say that the result happens because the future state is to be achieved. What this means is that the theoretical determination of the description that the future state will be achieved leads to the theoretical determination of the description that the result will occur. We may also say that the result happens *in order that* the future state be achieved. The person gets a drink in order that their thirst be satisfied.

Someone who still thinks that there is a problem here with backwards causation may have the confused thought that a real causal mechanism cannot be represented by a theoretical method that goes off into future descriptions before coming back to describing the result. The process of such a mechanism operating would seem to have to go forwards then backwards; but processes only go forwards. The thought is confused because there is no reason to suppose that the process of a mechanism operating must work in the same order as the process of the method of intellectual determination being applied.

Perhaps there is thought to be a problem with the idea that in different circumstances that part of the process of intellectual determination that results in the description of the result given the description of the future state could be altered. This means that the description of the future state would remain unaltered while the description of the preceding result alters. If the actual teleological process is represented by this, then the preceding result must be altered by circumstances that affect how that result leads to the future state. This looks like backwards causation again.

But of course it is not backwards causation. For, this is exactly the way teleological mechanisms do work. If something blocks the causal path from getting a drink to satisfying one's thirst (e.g. one's mouth is sewn up), then some other result like getting an intravenous injection of water will occur while the future state remains unaltered. All that is required to avoid backwards causation is that the circumstances that affect how the result leads to the future state should have come into being before the result is determined. Since the theoretical method only has access to information about how the world is before the result occurs, this is ensured.

The second kind of doubt there might be about the real causal status of teleological explanation is based on the fact that phenomena which are teleologically explained may also have perfectly good non-teleological explanations. Consider, for example, the so-called physicalist explanations of human activity in terms of nerve firings

etc. Does this not indicate that teleological explanation can have at best only pragmatic and not ontological significance?

One response would be to distinguish the sorts of things that are explained teleologically from the sorts of things explained non-teleologically. For example, the intransitive movement of my hand should be distinguished from my transitively moving my hand. The two phenomena are different and may be explained in quite different ways. (See Hornsby 1980 for a use of this distinction.)

This sort of distinction seems to be important in giving a basis for the distinction between belief-desire explanations and neurophysiological explanations. However, I do not want to assume that teleological explanation must be belief-desire explanation. In general, I see no reason to deny that intransitive as well as transitive body movements can be explained teleologically. So, what I have to show is that it is possible to explain the very same phenomenon with two distinct explanations both of which are causal.

The assumption according to which this is impossible is that nothing can be determined by each of two totally different mechanisms. Given this assumption and assuming that we reject any help from the response that teleological explanations explain distinctively teleological phenomena, one of two things follows. Either, teleological and non-teleological mechanisms are not distinct; the apparently distinctive theoretical method embodying practical rationality actually represents a mechanism that can be represented just as well by a theoretical method that does not embody practical rationality. Or,

teleological theoretical methods do not represent real mechanisms; in which case teleological explanation is not real causal explanation.

Either of these alternatives would be pretty grim. But fortunately, we can reject the assumption instead. For it *is* possible for one phenomenon to be determined individually by several distinct mechanisms. It *is* possible for one phenomenon to be the result of several different processes.

It should be clear from chapter 1 that I reject the ontological vision that we should describe the world in terms of a minimal set of independent entities. This rejection is made independently of the aim to make teleological a real kind of causal explanation. Take an example from physics. Suppose that a sample of gas is released in a room and then, after say 20 seconds a threshold number of gas molecules - say 10 - is registered by a sensor on the other side of the room. One mechanism operating here is a diffusion mechanism, the working of which depends on air pressure, temperature, currents, etc. Another mechanism is the set of ten kinetic mechanisms, one for each molecule in question, the workings of which depend on starting velocities of each gas molecule and positions and velocities of the particles.

Some phenomena can be explained just as well in terms of the working of either mechanism. In these cases both mechanisms are working. Some phenomena can only be explained in terms of the working of one, and some phenomena can only be explained in terms of the

working of the other mechanism. In these cases just one of the mechanisms is working.

Some people might assume that neither of these are real mechanisms; there must be a Super Mechanism that determines all the phenomena. But why should there be? And what is wrong with leaving these lesser mechanisms existing together? If you have ways of identifying both mechanisms, then this should mean that when the identification is successful in both cases, they both exist. Similarly, there are two different processes happening at the same time and both resulting in the same phenomenon.

The motive for ruling out this possibility might be the thought that there is something incoherent about causal overdetermination. I think that the thought may be quite correct but misplaced if used in this way. Let us distinguish different kinds of over-determination. There is the theoretical kind of overdetermination that is quite unproblematic. This is exemplified by the case of the lit match and the bolt of lightning theoretically overdetermining the description that the piece of paper will ignite, even though only the lightning causally determined the actual event of the piece of paper igniting.

Next, there is metaphysically independent causal overdetermination. This might be exemplified by the case of the two bullets piercing someone's heart simultaneously. In this case, two processes that are supposed to be quite independent of each other each result in the death of the victim. This is rightly thought to be incoherent or at least very unusual. If there really is no way of separating out one of

the processes as the *one* that resulted in the death, then what resulted in the death must be a conglomerate process involving both bullets. Since the death is actually part of the process of a bullet killing the victim, there cannot be two completely separate processes each resulting in the death.

There is a third sort of causal overdetermination, which is the one that I am interested in here. In this case, different processes that are not metaphysically independent of each other each results in the same phenomenon. There is nothing in the argument against the possibility of metaphysically independent overdetermination that applies also against this metaphysically interdependent causal overdetermination.

I suspect that the motive for ruling out the possibility that metaphysically interdependent mechanisms or processes exist together is that science is thought to be messed up by them if they are allowed to stay. Such mechanisms do not interact, so we cannot provide a single complete and coherent theoretical method representing them both together. I hope I showed in chapter 3 that this scientific ideal of having just one story to explain everything should not stop us from accepting as real the mechanisms that science in its present state helps us identify.

This view allowing a plurality of mechanisms is a special case of the widely held view allowing distinct but metaphysically interdependent objects. For example, a plant and the stem of a plant are not the same entity even when the plant has no other parts than that stem. We are assuming here that we have a rootless plant that

will one day flower. The flower will not be part of the stem but will be part of the plant. So the plant and the stem, which, before the flower emerges, may occupy exactly the same space, are nevertheless at all times distinct entities.

There is no question of these two entities competing with each other for space. Similarly, we should not feel that metaphysically interdependent causal processes or mechanisms have to compete with each other for causal space. Such entities are not compatible with each other for purposes of counting, interacting, competing for space, etc.

There is a feeling about science that ultimately it should provide us with a single way of picking out a fundamental class of entities that are all capable of interacting with each other; and then these will be the only entities that *really* exist. All other so-called entities will be metaphysically dependent on them. Without conceding anything to this view about realism, I could remain unruffled by its conclusion, if it was that teleological mechanisms are no more real than tables or oak trees. I should be more worried about an argument that showed that teleological mechanisms were less real.

So, the fact that phenomena that are teleologically explained may have perfectly good non-teleological explanations does not imply any ontological redundancy in teleological explanations. Nor, of course, does this fact show that there is any explanatory redundancy in teleological explanations. We can explain a piece of body movement teleologically when there is no other non-teleological explanation

available to us. Indeed a neurophysiological explanation would be quite useless to us even if it were available. What we are interested in is seeing how the body movement contributes to the goal-directed action of a rational agent. And we cannot do without the teleological explanation to see that.

The third kind of doubt there is about the causal status of teleological explanation is based on the fact that the theoretical method apparently representing how a teleological mechanism works only provides us with reasons why a phenomenon *should* be achieved, and does not provide us with reason why it *must* happen. It might seem that a teleological rationalization could not get to grips with the deterministic nature of the world and real causal processes. So it might be complained that there is something preposterous about the idea that practical rationality is the basis of a *causal* explanation since only theoretical rationality can have that role.

I think that this is only a forceful argument if something like the D-N account of causal explanation is assumed. It is not true (given my arguments in chapter 2) that in order to explain why something happened it is sufficient to show that it is theoretically rational. For, if it were sufficient, then the notion of theoretical rationality would have to be so strong as to be entirely impossible. So, there is no problem either with the fact that one might describe something as practically rational without that thing necessarily happening. In either case, there must be a mechanism whose working is represented by a theoretical method that embodies either the theoretical or the

practical rationality, and which results in the phenomenon to be explained.

Neither theoretical nor practical explanations ever do or need to provide us with the conclusion that some phenomenon *must* occur. All we may conclude is that that phenomenon *should* occur according to the theoretical method being applied. The mechanism is what actually *makes* the phenomenon occur.

4.3 Means-Ends

Central to any sort of objective practical rationality is some version of means-ends analysis. In this section I aim to establish the basic constraints that embodiment of means-ends analysis imposes on a theoretical method. I go on to show how the fact that means-ends analysis is involved in practical rationality and in teleological explanation accounts for much of the strength of teleological explanation. I.e. the basic constraints that means-ends analysis imposes are only satisfied by purposeful systems.

What sorts of things are means and ends, and how do they relate to each other? Let me start with the idea of a way of a state E becoming realized. This idea has no implication of agency about. For example, a way of the stone becoming hot is the stone being warmed by the sun. Or, a way of England becoming covered by cumulo-nimbus cloud is an

occluded front coming in from the Atlantic meeting a low pressure area moving across from central Europe.

A way of a state becoming realized is a *route* to that state being realized. A route is a series of processes leading to the result. So I suggest that a way is just a process or series of processes that results in the end state being realized.

It is clear that every process which results in some state becoming realized counts as a way of that state becoming realized. But, not every *series* of processes resulting in some state counts as a way of that state becoming realized. Only those processes that have a role in the final outcome should be included.

The converse also holds. Every way of a state becoming realized is a process or series of non-redundant processes leading to that state being realized. A way could not just be a series of events instead, because the way must *make* the end state become realized. The way of a stone becoming hot is not just the event of the stone becoming hot, nor is it any of the preceding events, nor all of them together. The way does not merely cause the result; it leads up to it.

It is the process of driving to Brixton, turning onto the A23 and driving to the end of that road that is the way of getting to Brighton. The events of arriving at Brixton and finding oneself on the A23 are not parts of the way of getting to Brighton. Processes rather than events make things happen.

Specifying a way is an anwswer to the question of *how* some state becomes realized. Similarly, one can ask *how* the way itself is

achieved. 'How is the stone being warmed by the sun? How does one drive to Brixton?' What is *not* being asked for here is a specification of how the underlying nature of the process becomes present. The way that the stone becomes warmed by the sun is the series of processes of radiation being emitted from the sun, this radiation travelling from the sun to the stone, the radiation being absorbed by the stone and the temperature of the stone rising through this absorption. The way one drives to Brixton is the series of processes of driving to The Elephant and Castle, turning on to the Kennington Park Road, etc. We do not describe as the way of driving to Brixton the process of becoming motivated to driving to Brixton.

It is worth mentioning that we do not describe driving to Brixton as the way of driving to Brixton. The way must be a breakdown of the process. It cannot just be the original process itself.

It may seem that because there can be a way of achieving a way, the end of a way need not be a state of affairs but may itself be a process. But a way of achieving a way is not a way of making a process be happening, but a way of making it the case that the process has happened. So, when one asks how does the way itself happen, one is just asking for a finer-grained structure of processes leading to the end result.

So a way is a process or series of processes, an end is a state of affairs and the relationship between them is that the way results in the end becoming achieved. The way itself may be further broken down into sub-processes. Note that this does not constitute a reduction of

a process into a structure of sub-processes. The same process might have entirely different structures of sub-processes.

I suggest that the notion of a *means* to an end can be understood in terms of that of a way of an end becoming realized. A means to an end is a sub-series of a way of the end being realized. The way is a series of possibly overlapping processes (each one of which must happen if the whole way is to be effective) resulting in the end, E, occurring. A means is a subseries of the way. For example, a means to a person getting to Brighton is that person getting to Brixton, etc.

(When we talk of *the* means rather than *a* means to an end, we usually mean the whole of the way rather than just part of it. But nothing much rests on this distinction.)

It seems that not just any way counts as a means. It is absurd to say that a means to a stone becoming hot is the sun emitting radiation. The means must be somehow controllable; it must potentially be an action.

If this is included as a constraint on what counts as a means, my entire account collapses. I am trying to provide an account of purposefulness and agency in terms of adaptability to means-ends analysis. If part of what it was for something to be a means was that it was potentially an action, then the account would be uselessly circular. However, the objection of circularity would fall away if it was a consequence of the right account of what a means was that it was potentially an action.

I suggest that there is no absolute way of determining what ways of achieving an end are to count as means. Rather, each particular system of practical rationality will have its own constraints on what processes may count as means. Some basic types of process will be allowed and then processes consisting of structures of these basic types will be allowed also. For example, one system of practical rationality might count a certain sort of movement of a robot arm as a basic means. Any process consisting of a series of such movements would also count as a means.

I make no assumption that the allowable means must be in the control of an agent. This will turn out to be a consequence of the means being available to a method of practical rationality that does explain something. A method of practical rationality that included as a possible means the process of lightning striking things would fail to represent any mechanism. So there is no agent (with the possible exception of Thor) for whom lightning striking would count as a means. If the system of practical rationality which included the robot arm movements as means did represent the functioning of some mechanism, then such movements would be potential actions.

The details of this account of agency will be provided later. All I want to show now is that I do not have to smuggle in a notion of agency in order to talk about means-ends analysis.

There are some intriguing questions about practical rationality that I intend to ignore for the time being. 1. What is it in a theoretical method embodying practical rationality that sets off the

whole chain of means and ends? In other words, how are top level goals established? 2. Must the means be the *best* way of achieving the end, or will any way that works count as a means? 3. Does practical rationality work top-down from ends to means only, or is it a looping method in which the possible means should condition the choice of ends? For the purposes of this chapter, I concentrate on top-down means-ends analysis as characteristic of practical rationality, and leave these questions to chapter 5.

So, what can we say about a theoretical method embodying practical rationality? First of all, it must have ways of describing certain states as desirable. Then it must provide descriptions of processes as desirable that lead to these desirable states being realized within the constraints of the particular practical rationality system. This part of the method regards some processes as fixed and certain others (the means) as variable.

Note that working such a theoretical method is not just reading things off a fixed theoretical structure. It involves finding out what will lead to what. A variant of such a method describes, not what will lead to what, but what is most likely to lead to what. But, of course, this is an empirical matter also.

In 1.4 I criticized standard externalist accounts of teleological explanation, like that of Charles Taylor for instance, for failing to capture the notion of practical justification.

> *To offer a teleological explanation of some event or class of events, e.g., the behaviour of some being, is, then, to account for it by laws in terms of which an event's*

> *occurring is held to be dependent on that event's being required for some end.*
> (Taylor 1964 p9)

That an event's occurring is held to be dependent on that event's being required for some end is not a sufficient condition for us to be able to say that the end is practically rational. Because this condition is too weak, it seems that the activity of water in a U-shaped tube counts as teleological on Taylor's account. I should now be in a position to say what it is about practical justification that is lacking from Taylor's account.

One suggestion is that the notion of evaluation is lacking. Means-ends analysis must involve evaluation of goals. Evaluation depends on essentially personal considerations and needs. So, the balancing of water levels in a U-shaped tube cannot be described as a desirable end.

Although this suggestion may be true, I object to its being part of the analysis. For one thing, it is rather brutally stipulative. But my main objection is that it requires the introduction of some notion of agency at too early a stage.

My suggestion is that what is lacking from Taylor's account is the requirement that a means to an end must be a process. This is a surprisingly important consideration. Without it, we would be able to think of the *event* of the water level in one of the arms rising as a means, separate from the end of the water levels becoming equalized, and resulting in it. Whenever this event is required for the end to be

achieved it occurs. So, it would look as if some sort of practical rationality was at work here.

However, when we see that a means must be a process, it is clear that we have not got a separate means to the end of the water levels becoming equalized. One way of the water levels becoming equal in the arms of the U-shaped tube is the process of the rising of the water level in one of the arms of the U-shaped tube under gravitational pressure. Gravity is an essential part of this process. Without it, any rising of the water level would have to belong to some other process altogether.

But now it is clear that nothing real is being represented by the move from the description that the water levels will equalize to the description that the water in one of the arms will rise under gravitational pressure. There is no sensitivity in any mechanism operating here to whether or not the process of the rising of the water level in one arm under gravitational pressure is the best way of equalizing the water levels. There are not two processes here, but only one.

The apparent element of teleology in this example arises from the fact that there is a mechanism operating here which is sensitive to whether for example a rise in the water level in one of the arms of two inches or of three inches say is required for the levels to be equalized. But although a rise of two inches and a rise of three inches are different events, there is only one process which they belong to. Whether or not that process happens does not depend on

what is required or what is the best way of achieving equalization of levels.

So, the rising of water in one of the arms of the U-shaped tube cannot count as a means to the equalization of levels, since it is the very process of the equalization of levels. Why then are we entitled to say that moving the white queen to B6 is a means of checkmating black when that move is the checkmating move? It might be argued that this example suggests that it is the description of the process that is important in separating means from ends rather than a separation in the processes themselves.

But this chess example is an example of separate processes after all. The process of moving the white queen to B6 is not the same as the process of checkmating black. The processes have the same constituent parts, they have the same result, but they are nevertheless distinct. The characteristic structures of events are distinct, but happen to coincide in this instance.

4.4 Are 'Smart' Systems Necessarily Teleological?

I shall argue that it is not a valid teleological explanation to say that a bird sits on its egg because this is the best way of achieving incubation of the egg. If another way of incubating the egg was better, the bird's activity would not adapt. It would still sit on its egg even if, say, leaving it on top of a radiator and going walkabout was a much better way of getting it incubated. This would be the case however good the evidence.

This is not simply a case of the bird's teleological mechanism failing to operate or being interfered with. For if such were the case, then it should be possible to rig up perfect conditions in which the failure of the mechanism did not occur and nothing interfered. But there is simply no way that one can persuade a bird to do what it should do, if what it should do is anything different from sitting on its egg.

But perhaps there is a teleological explanation to be found here not in the individual but at an evolutionary level. Can we say that development of an egg-sitting mechanism in a species of bird occurs because that leads to the incubation of the eggs of members of that species? Again no! This sounds OK partly because expression of Darwin's theory of evolution is often very careless. But it is a bogus explanation as I shall show in 4.5.

In this section I show that non-purposeful activity that may seem on first sight to be teleological is not. The activity of one's body

organs is not teleological; nor is purely instinctive activity of animals; nor is the programmed activity of present-day computer systems.

Let us start by considering the mechanism which maintains the blood sugar level in one's body within a certain range. Very roughly, the pancreas produces more or less insulin in response to the concentration of sugar in the blood stream, and the liver responds to the concentration of insulin by metabolizing more or less sugar. This mechanism is homeostatic and can loosely be described as directed to maintaining blood sugar level equilibrium. But, as I will show, the mechanism is not really teleological at all.

Suppose that a situation might arise where something other than the normal response of the pancreas is required to keep blood sugar at the correct level. Perhaps a massive secretion of insulin is required right away to cope with the injection of a strong glucose solution into your blood stream that will be made in a few minutes. Your pancreas will respond to the injection after it is made, but this will be too late. The only way to stop the blood sugar level from rising to a fatal level is for the pancreas to start working now, which of course it will not do.

Now there is no requirement that mechanisms be 100% reliable. Things may interfere with the structure of a mechanism so that it fails to operate and the characteristic results fail to materialize. So, there is no problem with the thought that the characteristic result of the mechanism controlling the blood sugar level fails to materialize

because there is something lacking in the mechanism in this case. The operational conditions which are part of the mechanism are not satisfied. If they were satisfied, then the mechanism would result in the blood sugar level being in the correct range.

But it is a very different story if we consider the question of whether it is a characteristic result of this mechanism that the *way* of achieving the correct blood sugar level is achieved. In the situation being considered, the mechanism fails both to achieve the correct blood sugar level and to achieve the way of achieving the correct blood sugar level, which in this case is the early secretion of insulin. Now, although the operating conditions can be improved so that the mechanism does achieve the correct blood sugar level, there is no way of improving the conditions so that it achieves the way of achieving the correct blood sugar level, i.e. the early secretion of insulin. The failure to achieve the way to achieve the correct blood sugar level cannot be attributed to a failure of the mechanism or else there would be a way of remedying that failure so that the mechanism works properly and achieves that result; but there is none.

Contrast this with the parallel case where human purpose is involved. Consider the mechanism of a diabetic achieving the best way of keeping the correct blood sugar level. We can construct similar examples where they fail to achieve the best way. But always it will be possible to attribute such failure to a lack in the operational conditions of the mechanism. For, in those circumstances where it is apparent to the diabetic agent that some activity, A, is the best way

of achieving the correct blood sugar level and their brain is working properly, then A will be done. The control mechanism of the diabetic agent may be much less reliable than the body's natural mechanism, but at least it can be described in terms of means-ends analysis.

Take another example. There is a family of computer systems known variously and misleadingly as expert systems, intelligent knowledge-based systems. etc. These systems may include sensory input devices and robotic output devices. They are programmed with a series of 'rules' or 'facts'. They can produce the logical consequences of these 'rules' or 'facts', and thus condition their activity to their sensory inputs according to these 'rules' or 'facts'. Usually these 'rules' or 'facts' are supposed to encapsulate human expertise or knowledge in the particular tasks these systems are designed to perform.

The activity of such systems is often optimistically described as smart, intelligent or rational. Certainly it may seem that way on superficial examination. However, the activity of such systems is never really teleological. Even if the results often turn out to be what is rational, rationality is not *characteristic* of such systems. The way to achieve an end can never be finitely specified in the way required by such systems.

The pattern of inputs the machine reacts to may be a reliable indicator of the actual situation in which the activity the machine happens to be programmed to respond with is the way of achieving some end. But the pattern of inputs is distinct from this situation. There can always be devised some situation which the pattern of inputs fails

to pick up properly. In such a situation the appropriate activity is not achieved, and there is no way of improving the operating conditions so that it is achieved.

The only way to get the machine to come out with the right activity in such a situation is to change its program. But changing the program cannot be described as establishing optimal operating conditions for the mechanism. The machine with the old program operated as a different mechanism to the machine with the new program. So even when it had results which were rational in some sense, it was not characteristic of the way it worked that its results were rational. Indeed it is not characteristic of the machine with the new program that its results are rational either. However many extra 'rules' or 'facts' are added to the program, situations can always be constructed in which the machine, however well it is working, will inevitably fail to produce activity that is the best way of achieving whatever end the machine has been designed to achieve.

(Now, if we include the computer programmer adapting the machine to cope with new types of situation as part of the mechanism we do have a practical rationality mechanism. But the mechanism is only practically rational in virtue of the programmer's continuing presence as part of it. Left to itself, the machine is not teleological.)

A natural response to this might be to ask why we cannot include as part of the mechanism's operational conditions that the best activity in the circumstances is the programmed activity. If this condition is included as part of what it is for the mechanism to be

operating, then we cannot devise situations in which the mechanism is operating but the supposed characteristic result (that is the best means to achieve the end) is not achieved.

At first sight, this looks to be a perfectly acceptable move. Suppose our expert system has been designed to maintain some pattern of temperature levels in the rooms of a large building. And suppose that some situation arises that the machine has not been programmed to cope with, for example a window is left open beside one of the temperature monitors. Because of this, the best thing for the system to do is not done, and the desirable temperature pattern is not achieved.

It is an aspect of my account of mechanisms that the possibility of this situation does not mean that it is not a characteristic result of the mechanism when it is working that it maintains such-and-such a temperature pattern in the building. For, the mechanism may be considered as including in its operational conditions aspects of the world outside the actual machine. The machine is only part of what makes the result; the environment of the machine has a part to play in the process as well. So, we can say that when the window is open, something essential to the process of maintaining the temperature pattern is lacking, and the mechanism, properly described as characteristically resulting in this temperature pattern being maintained, is not operating.

Now, why cannot this account be applied to show that the mechanism can properly be described as characteristically resulting in the *best*

means to achieve the maintainance of the temperature pattern? The answer is that by stipulating that those situations in which the program does not result in what is best count as situations in which the mechanism is not operating, we are stipulating out of existence part of the supposed characteristic result.

This can be seen clearly using a simple example. Suppose that I have programmed a computer to work in the following way. Given any input of an integer in the range 0 to 10, that same integer will be output unless the input is '2' in which case there is no output; for any other input (including no input at all) nothing will be output. Would the following describe a mechanism at work here? 'Given any input of an integer in the range 0 to 10 that same integer is output; for any other input there is no output.' In other words, can we include among the operational conditions of the mechanism the fact that '2' is not input along with other requirements like the machine being switched on, etc?

The answer is no. The theoretical method that is supposed to represent the way the mechanism works discriminates between situations in which different integers are input. The method represents the mechanism only if there is a way of the mechanism working (the essential nature of the mechanism working includes its operational conditions being present) which results in what is described by the theoretical method in all the types of situation discriminated by that method. But, there is no way of the mechanism working - no set of

operational conditions - where in addition to outputting all the other results described by the method, it also outputs '2' when '2' is input.

So a theoretical method that represents a mechanism cannot specifically include a possible kind of situation that is then excluded by the operational conditions of the mechanism. For this reason, a theoretical method incorporating practical rationality (or, as we are simplifying in this chapter, means-ends analysis) does not represent the way the heating expert system works. A theoretical method incorporating means-ends analysis will discriminate between an indefinite number of types of situation relevant to what is the best way to achieve the end. The expert system, in its working, is not sensitive to any more than a finite subset of these differences. We are not entitled to say that the mechanism is just not working when any of these other situations arise. What we must do instead is to limit the scope of the theoretical method so that it discriminates between only those different situations which the mechanism is sensitive to. A theoretical method so reduced will no longer incorporate means-ends analysis.

One final attempt to see the expert system as teleological might be to adapt the notion of desirability to the machine's limited capacities. Something would only count as desirable if the system could come to the conclusion that it was desirable. Presumably, this would just involve the system having an internal 'representation' of the means connected up on one side by bits of logic programming to an internal

'representation' of the end, and on the other side to some means-achieving output.

This view is a version of an internalist account of teleology involving a subjective kind of practical rationality. It will be very much my concern to show that such an account is wrong. But this will be done in chapter 5 when I consider what we mean by practical rationality. For the time being, I will simply stipulate that the kind of practical rationality required in an account of real teleology must be objective. Something can seem to a person to be the best way of achieving an end and it not really (objectively) be the best way. Similarly, a machine may have the appropriate sort of internal state linked to a means-achieving output and yet that means not really be the best way to achieve the end.

Indeed, I suspect that all this talk of machines (or anything else) having internal representations depends for its cogency on the machines working teleologically. I do not think that we can help ourselves to the notion of internal representations and then use this notion to introduce teleology in an internalist way.

None of this is to say that machines cannot be teleological. The current boom in parallel distributed processing (see Rumelhart and McLelland 1986) provides us with machines which adapt their programs in the light of failures and successes. They are still not teleological, because they always react to past failures rather than working out in advance what is best. But, if you combine these with perceptual devices that actively test hypotheses in the process of

gleaning information as well as internal trial and error processes, I think teleology will start to seem attainable. Such confidence is supported by the fact that *we* can act purposefully, and yet a person can be made by sticking together a load of neurons and connecting them up to perceptual organs, life maintaining organs, and action organs. Do the same thing but with stainless steel, and you have a baby teleological robot.

Now, there may be some worry that my account of teleology is so strong that even people cannot be said to act teleologically on this account. It may seem as if my account requires that people be omniscient. And, all sorts of other doubts may still remain about how real practical rationality mechanisms can possibly work. I hope to deal with these in the next chapter. I hope that I have at least shown here that if real practical rationality mechanisms can work, then present day pancreases, computerized thermostats or incubating seagulls do not embody them.

4.5 Functions and Purposes

Given my grand stategy, that of taking an account of teleology all the way to being an account of purpose in one stride, it is important to show that non-purposeful functions are not teleological in my sense. At the same time, it is important to account for our tendency to throw evolutionary functions in with teleology and real purpose.

The aim of this section is to characterize a distinction between two kinds of teleological explanation, explanation in terms of purpose and explanation in terms of function. The distinction can be illustrated by these examples

"He put on his hat in order to look smart."

"The hat has got a brim so that rain falling on the hat will not drip down the neck of the wearer."

The first explanation is a purpose-giving explanation. The activity of the person is explained in terms of what is likely to follow from that activity being performed by the person. The second explanation is a function-giving explanation. A feature of the hat is explained in terms of what is likely to result from the hat having that feature. The two kinds of explanation seem to be closely related. The most obvious difference between them is that in the first case a person's mental states are essentially involved, and in the second case the hat's mental states do not enter into the story at all. On the other hand, it might be claimed that the mental states of the *designer*

of the hat rather than of the hat itself *are* essentially involved in the second explanation.

The central kind of teleological explanation for an internalist is purpose-giving explanation. The natural way to move from this to an account of function-giving explanation is to regard such explanation as teleological in virtue of there being some purpose-giving explanation holding in the situation. If the presence of a feature, like the brim of a hat, is the result of activity that can be explained in terms of purpose, then that feature can be explained at one remove in terms of that purpose. So, the brim on a hat is the result of the activity of some prehistoric hat designer who had the purpose of keeping water from running down the wearer's neck. According to an internalist, the hat brim has that function because the activity resulting in the hat brim had that purpose.

This reduction of function-giving explanation to purpose-giving explanation breaks down however in the case of evolutionary or natural functions. For example, the function of chlorophyll in a plant is to enable that plant to photosynthesize. But, the presence of chlorophyll in the plant is not the result of any activity that can be explained in terms of purpose (at least in the internalist sense of involving future-directed mental states).

The internalist must respond in one of two ways.

1. They may concede that there is no such thing as natural function; that the language of natural functions developed in the days when people believed in the role of divine purpose in evolution; but that such language is now defunct or at best metaphorical.

2. They may claim that function-giving explanation is not so directly related to purpose-giving explanation after all. So, for example, what we mean when we say that something has a function is that it is *as if* it is present for a purpose.

An *externalist* however claims that we can make sense of teleological explanations without reference to future-directed mental states. A formal feature of a teleological explanation is that something is explained in terms of its likely results. Externalists show how we can attribute such explanations on the basis of the adaptability of a system to certain requirements without reference to that system's mental states. As a result, externalists tend not to make a significant distinction between purpose-giving and function-giving explanations. There is with most externalists and internalists alike an underlying assumption that if we do not base the distinction in something substantial, like whether or not mental states are involved, there is no significant distinction to be made.

Consider an evolutionary case of a function-giving explanation. Is there anything wrong with an explanation like this, 'The reason we have hearts is so that oxygenated blood gets pumped round our bodies?' If there is nothing wrong with this as an explanation, isn't it a classic case of teleological explanation? Aren't we explaining something in terms of its being a means to an end? And if it is a teleological explanation (and assuming of course that no real purpose is involved in our acquiring hearts), doesn't this mean that teleology does not characterize purposefulness?

My strategy is to show how in all cases of function-giving explanation the phenomenon is not in fact explained in terms of its being a means to an end. In such explanations there is something which is a means to an end, and that fact is an essential part of the explanation of the phenomenon. But it is not the phenomenon itself that is the means to the end, so the explanation is not truly purpose-giving.

If this strategy works, then I can hold onto the idea that if something happens because its happening is a means to an end, then it is not only teleologically explained, but it is purposeful. So, I can maintain an incredibly simple account of purpose without having to beg the question by insisting on some intentional notion of what a means to an end is.

There are two ways in which people may try to attribute goals to non-purposeful systems, and they must be carefully distinguished. Take for instance the blood sugar equilibrating mechanism in someone's body. First of all, its presence in the bodies of members of our species may be explained in terms of the functioning of that mechanism being a means to some end - survival for instance. Secondly, the activity of that mechanism may be explained in terms of that activity being a means to an end - the blood sugar level being maintained at the correct level for instance.

In the first case the supposedly teleological mechanism is the result of evolution. In the second case, the supposedly teleological mechanism is the mechanism of evolution itself. I have argued in 4.4

that the first sort of explanation is not teleological at all. It is simply bogus to introduce talk of practical rationality and goals in describing how such mechanisms work. In this section I shall argue that the second sort of explanation, if carefully expressed is perfectly correct. My aim is to show why it is not the same as purpose-giving explanation.

So although the activity of the pancreas is not to be explained in terms of purpose, we will see that there is some teleological aspect to the pancreas. It is not the activity which comes from the pancreas, but the activity which produces the pancreas that depends on practical rationality.

In considering this, it is really a red herring that the systems in question partially adapt their activity to the requirements of the situation. For, the teleological aspect of the pancreas is like the teleological aspect of such passive entities as a corkscrew or the colour on a butterfly's wings for instance. All these things have functions. Their having functions is not an aspect of explanations of their activity; it is an aspect of explanations of their presence. The process that results in their presence is what is teleological in these cases. Processes that result *from* their presence need not be.

The accounts of purpose-giving and function-giving teleological explanation can now be made explicit.

Activity, X, is teleologically explained in terms of purpose, Y, if and only if X is the result of a mechanism which characteristically results in activity which is described as rational by a theoretical method embodying objective practical rationality; and the fact that X is the

means to the achievment of Y is essential to X's being described as rational according to that method.

Feature, A, is teleologically explained in terms of function, B, if and only if the presence of A is the result of a mechanism which characteristically results in the presence of a feature in such a way that the presence of this feature is described as rational by a theoretical method embodying practical rationality; and the fact that the functioning of feature A is the means to the achievment of B is essential to the functioning of feature A being described as rational according to that method.

So, an explanation of purpose-giving form explains some activity in terms of its being rational (a means to an end) according to some version of practical rationality. The activity happens because it should happen. Whereas, an explanation of function-giving form explains the presence of some feature in terms of the presence of that feature being rational (a means to an end) according to some version of objective practical rationality. The feature is present because the feature is useful.

Going back to the first example, the activity of a person putting on a hat is the result of a mechanism (the person's mechanism for acting) which characteristically results in activity that is rational according to some version of practical rationality. That putting on a hat is a means to the end of looking smart is essential to the fact that putting on the hat is rational according to that version. The activity is explained in terms of its being rational. This is why it is purposeful on my account.

The presence of a brim on the hat is the result of a process (the hat design and development process) which characteristically results in

the presence of hat features that are rational according to some version of practical rationality. That a hat having a brim is a means to the end of keeping the rain from dripping down the wearer's neck is essential to the fact that the presence of the brim is rational according to that version. The presence of the brim is explained in terms of the rationality of having a brim. This is why it has a function on my account.

Now, it may seem that on this account, an explanation of function-giving form must also have purpose-giving form. Consider the following stages of the argument for this collapse.

1. Assume that there is a function-giving explanation of the presence of a feature A - say the brim on a hat.

2. Then, there is a mechanism resulting in A's becoming present which is sensitive to whether or not A's presence is practically rational.

3. This means that the mechanism resulting in A becoming present is sensitive to whether or not A becoming present is practically rational.

4. The mechanism resulting in the process of A becoming present is sensitive to whether or not that process is practically rational.

5. So, there is a purpose-giving explanation of the process of A becoming present.

This argument seems obvious to the point of being deductive. If a feature is present because it is useful, then the process resulting in the presence of the feature happens because it is rational that it should. But there is a very subtle false step here between stage 3 and stage 4. For this to be a valid step there must be an implicit assumption that there could only be one process of A becoming present.

In the interesting cases of evolutionary function, which I shall consider shortly, that assumption is false.

To see how this assumption might be false, consider the distinction between developmental and selectional explanations. (The terminology derives from Lewontin 1983 and Sober 1984.) Suppose the police are running an identity parade to get a positive identification on a suspect mugger. The witness' description includes the fact that the mugger was about six foot tall; so all the men in the identity parade have to be six foot tall. The story about police procedure helps to explain why the men in the parade are about six foot tall. It can explain why, if any man is picked out from the parade, he will be about six foot tall. But, it cannot explain for any man who is picked out why he is six foot tall. That can only be explained by some story involving his genes and his diet, etc.

The first kind of explanation can be described as a selectional explanation, and the second kind as a developmental explanation. It is important to realize that these two kinds of explanation are not just different ways of explaining the same phenomenon. Although it may be possible to explain the same phenomenon using both kinds of explanation, the basic difference between them is a difference in what is being explained. How Mr Robert Jones comes to be six foot tall cannot be explained at all in terms of police procedure . How the identity parade comes to contain men who are about six foot tall cannot be explained at all in terms of the stories about the genes, diet, etc, of that particular group of men.

Why we might think that there is only one thing being explained is that there is only one state of affairs of a six foot tall man, Mr Robert Jones, being present in the parade. But there are at least two processes resulting in that state of affairs - the process of Mr Robert Jones being selected for the parade and the process of Mr Robert Jones becoming six foot tall. And there is a different explanation of each process resulting in the presence in the parade of a six foot tall man, Mr Robert Jones.

It might be thought that Robert Jones coming to be six foot tall can be explained selectionally in virtue of the selectional explanation of how the identity parade came to contain six foot tall men and the fact that Robert Jones is in the identity parade. If explanations were entirely epistemic, then this would be the case. But, an explanation of a phenomenon is correct only in virtue of the existence of an actual process resulting in that phenomenon. The process resulting in Robert Jones being six foot tall has nothing to do with police procedure, even though one can infer that Robert Jones is six foot tall from knowledge of police procedure and knowledge that he was in the identity parade.

So we can see that there might be two quite different processes resulting in a feature being present. The mechanisms underlying each process will be sensitive to different things. We can see why the step from stage 3 to stage 4 is invalid now. The mechanism resulting in one of the processes of six-footedness becoming present might be sensitive to whether or not the other process of six-footedness becoming present is practically rational.

Now, apply this distinction to evolution. The presence of the feature of black colouration in a population of speckled moths can be explained selectionally in terms of the rationality of that feature. The evolutionary process resulting in the feature becoming present is sensitive to whether or not the presence of black colouration is practically rational according to a version of practical rationality. In this version of practical rationality, having a feature is rational if it helps the survival and propogation chances of an individual which possesses the feature. It is essential to the rationality of having black colouration in this version of practical rationality that getting black colouration is a means to the end of avoiding predation on pollution-affected trees. So, on my account, the function of the black colouration is to avoid predation on pollution-affected trees.

We can move down the steps of the argument above to stage 3. The mechanism resulting in black colouration becoming present is sensitive to whether or not black colouration becoming present is practically rational. But the mechanism is sensitive to whether or not the development of black colouration in an individual is practically rational not to whether or not the selection of black colouration in a population is practically rational. So one process is being explained in terms of the practical rationality of another process, and we do not automatically have a purpose-giving explanation.

(The point that evolution does not explain why an individual has a particular feature is made by Dretske 1988, pp 92,93. He uses this fact to distinguish real content from natural representation.)

It might be thought that there must be other ways of framing the evolutionary explanation in which one process *is* explained in terms of its own practical rationality. So consider some attempts to find such a way.

> *1. The process of a moth becoming black is explained by the fact that that process is practically rational, and, in particular, is a means of avoiding predation on pollution-affected trees.*

The explanation of the process of the moth becoming black must be regarded as a developmental explanation. But then it is clear that avoiding predation has nothing to do with it. A moth would still develop black colouration even if the environment had changed in such a way that this was completely irrational as a way of avoiding predation (e.g. all the pollution-affected trees have turned white). The development of black colouration in a moth depends in fact on its genetic material. It is quite unadaptive to whether or not it is rational for the process to occur in the circumstances.

> *2. The process of black colouration being selected in a population is explained by the fact that this process is practically rational, and, in particular, is a way of achieving a population, the members of which avoid predation on pollution-affected trees.*

Here it is true that the selection of black colouration is a way of achieving that end. However, the process resulting in this selection is not in fact sensitive to whether or not this is so. The process is

sensitive to whether an individual which develops black colouration avoids predation as a result, but this is a different thing.

There may be a confusion here due to the fact that the selection process has two separate parts at least. Let us say there is a first part which is a creation process through mutation and a second part which is a natural selection process. The natural selection of the feature of black colouration is explained in terms of the fact that the creation by mutation of that feature in an individual is a way of that individual avoiding predation. If these two processes are conflated, we may be led to the confused thought that one process is explained by *its* being a means to an end.

It is not a valid evolutionary explanation to say that a feature is present in a species because it is good for the species that that feature is present. This has been most convincingly demonstrated by Dawkins (1976). However, the nature of Dawkins' argument might provoke a worry that it is a contingent fact that considerations about survival of the species do not enter into evolutionary explanations. The opposing theory to selfish gene theory is group selection theory (see Wynne-Edwards 1962), and there is no deep theoretical reason why we could not explain selection of features by this theory. Would a group selection explanation have a purpose-giving form?

A group selection explanation has the following form:

> *3. The process of a feature being selected in a population is explained by the fact that the acquisition of that feature by a group of the population is practically rational for that group (i.e. leads to its survival).*

So, even a group selection explanation does not explain one process in terms of its own practical rationality. It breaks up the evolutionary process into three stages: initial mutation; development in a group of the feature (presumably by pure luck); selection of groups with that feature (resulting in the feature becoming present throughout the whole population). The third stage of the process may be explained in terms of the practical rationality of the second stage, but not in terms of its own practical rationality.

It should be obvious now why function-giving explanations do not necessarily collapse into purpose-giving explanations. For there to be a purpose-giving explanation of a process, P, we require that there is a mechanism that determines P, and is sensitive to whether P itself is the best way of achieving some end. This mechanism must be able to work in the light of what is likely to happen in the future even when this diverges from what has been happening in the past. This is not something that the evolutionary mechanism can do despite appearances to the contrary. Only rational agents determine processes in a way that is sensitive to what is likely to result from those processes.

Chapter 5

Practical Rationality

5.1 Humean Practical Rationality

The principle that I am defending is that a mechanism is teleological if and only if it characteristically results in what is practically rational. This principle can easily be accepted by internalists if they hold a subjectivist or Humean view of what counts as practical rationality. According to this view, something is practically rational only if it is *desired* or *intended* in a way that somehow overrides other desires (or intentions).

According to this Humean picture, there are primary desires and also beliefs about how those desires can be satisfied. The process of practical reasoning takes these desires and beliefs and produces new desires for those things which are believed to satisfy the original desires. The process continues and desires proliferate. The best or practically rational things to do are those things that are desired. If there is a conflict of desires, then the process of practical reasoning must work through this before the best thing to do is determined.

Combining this Humean account of practical rationality with the principle that a mechanism is teleological if and only if it characteristically results in what is practically rational gives us an

internalist theory of teleology. According to this theory, activity is teleological in virtue of being the result of a mechanism which adapts to the presence of the appropriate desires and beliefs. What desires and beliefs an individual is said to have are determined independently of characterizing the individual's activity as teleological. Then, teleological explanation of the individual's activity can be ascribed on the basis of the causal relation between these desires and the activity.

Although I am concerned to deny this account of teleology, it is worth mentioning in passing how this account is already an improvement over the standard causal account of action. The standard causal account required just that the activity be *caused* by the appropriate desire. Such an account is bedevilled by the problem of deviant causal chains (where unintentional activity is caused in a deviant way by the appropriate intention). But deviant causal chains are not a problem with this new improved internalist account where we require instead that the activity be the result of a process that characteristically involves desires being satisfied. If the climber drops the rope out of nervousness as a result of forming the intention to drop the rope and kill their friend on the other end of it, this is *not* determined by the process that characteristically involves putting intentions into practice.

By framing the account in terms of *processes*, the problem of deviant causal chains disappears. For, nothing can be the result of a process but only accidentally. The question of what it is for something to be caused non-deviantly is now replaced by the question

of what it is for something to be the result of a particular process. This is no longer a question in the philosophy of mind; the correct account of causation should be able to answer it.

So, if the internalist accepts my account of teleological explanation along with the free gift of a solution to the dilemma of deviant causal chains, then the distinction between this version of internalism and my version of externalism boils down to this: According to internalism, activity is teleological in virtue of resulting from a Humean practical rationality process. According to externalism, activity is teleological in virtue of resulting from an objective practical rationality process.

In 5.2 I consider what objective practical rationality is. Until then we can simplify the issue between Humean and objective practical rationality in the following way. According to Humean practical rationality, it is the *belief* that something is the best way of achieving the goal that is relevant to whether or not it is practically rational. According to objective practical rationality, it is the *fact* that something is the best way of achieving the goal that is relevant to whether or not it is practically rational.

As I mentioned at the beginning of 1.2, one can argue that Humean subjective practical rationality does not *really* justify action. All it provides is belief that an action is justified. Believing that something is justified does not mean that it is.

When a belief is rational, albeit false, then it may justify action in some sense. But this may be because, in order for the belief to be

rational, there must be some relevant facts behind the belief, and these make the action objectively justified in some sense. But now, if Humean practical rationality *did* justify action, then even totally (objectively) irrational beliefs would justify.

Suppose I believe that my car is parked in Street A. Suppose this is just some mental confusion and I really parked it in Street B but have temporarily forgotten. Is my action of walking up and down Street A justified by this belief?

I suggest that if one is inclined to answer yes to this question, one is forced to justify the answer by finding some objective facts to do the real justifying work. For example, perhaps I usually leave the car on Street A, so it is a good bet that I will do so on this occasion. But if this was the justification, then the belief is not relevant; the objective fact is doing the work.

Suppose I have never parked on Street A in my life and that the psychological reason for my belief is that I have been listening to someone talking about Street A. One is less inclined to say that my behaviour is justified. However, in even the most apparently irrational behaviour we can find some objective justification. In this case, there *is* an objective fact, which is that Street A has some significance for me. So on this basis there *is* some objective justification for believing that I parked the car there.

Even without any such flimsy justification for the belief, the action may be objectively justified given the belief. It is an objective fact that my unthinking beliefs usually turn out right; so it is a good

bet to act on this one. Note that it is not the belief that has any intrinsic justifying role here. If I was prone to getting voices in my ear saying, 'You have parked in Street X', and they usually turned out right, then it would be rational to go to Street X whether or not I had any beliefs about it.

If someone's unreasoned beliefs were generally a bad indicator of the way the world was, then that person would not be justified in following them. If this is right, Humean practical rationality is not really practical rationality at all.

Some theories of belief rule out the possibility of entirely irrational beliefs that are entirely unreliable (e.g. Dretske 1988). According to such theories, a belief is essentially a reliable indicator of the world. So beliefs will always justify action. But, of course, these theories accept that the justifying work is done by the objective facts underpinning the beliefs.

It may be thought that although Humean practical rationality only justifies in virtue of some underlying objective practical rationality, nevertheless actions can only be justified by Humean practical rationality; objective practical rationality by itself is not enough. The argument for this is that an objective fact cannot justify an action unless it is believed. Objective facts have to have some subjective realization as well if they are to justify action.

There are two possible intuitions behind this thought. One is that no fact can motivate an action unless it is believed. This is probably true, but does not support the claim that no fact can *justify* an action

unless it is believed. Given my account of teleological explanation, it is clear that these two questions should be kept separate.

The other intuition is that facts which are entirely inaccessible to an agent cannot be relevant to justification of the agent's activity. Only those facts which an agent has some access to can be part of a justification. From this, it is supposed to follow that only beliefs are relevant in justifying action.

There seem to be at least two ways of objectively justifying action (see 5.3 for more on this). One way is quite brutally externalist According to this way, you should do the action that has the right result even when you could have no way of knowing that that action would have the right result. According to the other way, only information that is available to the agent justifies any thing the agent does.

Now I see nothing theoretically wrong with the brutally externalist form of justification. But, even assuming that it is wrong and that only the second form of justification is acceptable, this justification does not collapse into Humean practical rationality. There is a distinction between information being available and information being believed. If there were no such distinction, then there would be no room for the underlying objective practical rationality at all. So, given all this, there is no reason to think that some special justificatory role is played by beliefs.

However, the case for externalism is not quite won yet. There are two kinds of objection an internalist can mount against the objectivity

of practical rationality. The first is based on the thought that we have no notion of what objective practical rationality could be. Quite apart from the question of whether we ever behave with objective practical rationality, there is the argument that there is nothing that would count as behaviour corresponding with objective practical rationality, because practical rationality cannot be objective. This is what I consider in 5.2.

The second sceptical attack is the argument that even granting we can make sense of what practical rationality would be, it is impossible that anybody's behaviour ever adapts to it. According to this argument, we are essentially subjective agents acting on information received and incapable of adapting to the way the world is beyond our beliefs and desires; so explanation of purposeful activity must be in terms of internal mental states and not in terms of the world outside. I consider this in 5.3 and 5.4.

5.2 The Possibility of Objective Practical Rationality

There is no difficulty in seeing how a theoretical method embodying objective practical rationality can incorporate goals. If the method yields the description that G is desirable or should be achieved, then G is a goal for that method. There are three ways in which such a description may be produced. It may be derived from some other goal by means-ends analysis (or some other more sophisticated decision

theoretic method). It may be a fixed aspect of the theoretical method - rather like an axiom. Or it may be derived somehow from the way the world is - neither a fixed axiom nor dependent on other goals. Having denied that there need be anything deductive about a theoretical method, there is no problem about deriving evaluative descriptions directly from consideration of the world.

The simplest sort of theoretical method would incorporate a single goal as an axiom, and work down from there by means-ends analysis. For example, consider the theoretical method that describes the activity of a company as practically rational when it leads to maximising the company's profits. Applying this theoretical method will result in varying descriptions of what activity is practically rational depending on the circumstances that apply. What will not vary is the description that profit maximization is practically rational, for that is axiomatic to the theoretical method.

Whether this theoretical method actually represents how a company works is not the present concern. The present claim is that there is nothing incoherent about a theoretical method that embodies practical rationality in the sense that it incorporates means-ends analysis. So there is no problem with objective practical rationality in theory. The real question is whether the more sophisticated kind of objective practical rationality that would be appropriate to teleological explanation of human behaviour is theoretically possible.

The first complication is that usually several goals need to be satisfied simultaneously. The way one goal is achieved is not

independent of what is required for another goal to be achieved. So means-ends analysis must deal with several goals together. Given this, the image of means-ends analysis being a single process from end to means is too simple. We should think of it as a looping process from ends to possible means and back again, continuing until a course of action is arrived at which satisfies a maximal set of ends.

This solution raises the possibility of a further complication. Need there be a unique result of this looping process of determining means and ends? What if the result was different depending on different ways of implementing the theoretical method? Indeed, this possibility of alternative means to an end is regarded as absolutely essential to practical rationality by Kenny (1975), for example. He argues that practical rationality works according to the principle of satisfactoriness. Practical rationality may determine several courses of action as satisfactory but not decide between them. A similar thought is that there is something creative about practical rationality, and that it does not follow a pre-set logic.

There are three possible ways of dealing with this. One way is to insist that the theoretical method has what it takes to distinguish between alternatives. This would force some arbitrary decision to Burridan's Ass situations. The problem with this is that if Burridan's Ass were to make different arbitrary decisions on different occasions, it would have to be seen as operating by different theoretical methods.

A second way is to allow imprecise results of a theoretical method. The method would determine a non-arbitrary description of what

Burridan's Ass should do; viz, it should go left or right. This seems to be a rather messy solution to the problem. It looks as if it would get out of hand with a course of action involving several of these alternatives.

A third way allows that a theoretical method may involve indeterminacy. If what I said at the end of 3.3 is right, then a theoretical method working according to satisfactoriness can represent a process with indeterminate results. Do we not often act with a sense of indifference, where there really is no reason for doing one thing rather than another? In this case an indeterminateness in the theoretical method embodying practical rationality is positively essential if it is to represent the way things actually work.

Another possible complication to this picture is that perhaps a theoretical method embodying objective practical rationality should still be sensitive to changes in superficial desires. For example, if I feel like having a cup of coffee, then it may be right for me to have one for that reason alone. Ruling out all subjective considerations in explaining behaviour rationally is going to give a false picture of how people operate. It will give the picture of people as unemotional reasoning machines; and this is a false picture.

One must be very clear about the following point. There are some states, like intentions, that cannot be ascribed to a person unless that person does have reason to act accordingly. I am arguing that it is a mistake to suppose that such states feed into practical rationality, at the bottom level anyway. There may be other states, like feeling like

having a cup of coffee, where it may or may not be the case that the person thereby has a reason for acting accordingly. A person may be operating according to a conception of practical rationality in which they should deny themselves what they feel like having. Or they may operate according to a conception of practical rationality in which they should satisfy what they feel like having if nothing very bad follows. In other words, a feeling may be a factor among many others which a theoretical method takes into account.

It might be objected that this puts strain on my aim in this thesis of showing how mental states are to be ascribed to a person on the basis of how their behaviour adapts to objective practical rationality. One response I might make is to accept that not all mental states are to be accounted for in this way. The strength of the account only depends on the propositional attitudes of intention and belief being ascribable in this way.

But there is an alternative response which maintains that it is only in virtue of how one can characterize a person as an objectively rational agent that feelings, etc as well as beliefs and intentions can be ascribed to them. The fact that these mental states may have a role in determining what is practically rational for an agent to do does not undermine the account so long as enough of what is practically rational can be established without them. In particular, the practical rationality on the basis of which these mental states are ascribed must not take such states into account.

All this talk is rather vague in the absence of a positive account yet of how mental states are ascribed on the basis of how a person adapts to objective practical rationality. Until such an account is produced in chapter 6 we can at least see that the objection that mental states may feed into practical rationality is not decisive against my account. As long as practical rationality can work at some level without reference to mental states then mental states can be ascribed on this basis. These mental states can then be fed into practical rationality at another level. The fact that I operate in a certain way is a factor that can be considered along with other factors in determining what it is rational for me to do.

There are many other complications that can be introduced into a theoretical method of objective practical rationality. The use of probability is one such. Simple means-ends analysis would need to be exchanged for some rather more complex decision-theoretic analysis involving risk analysis, etc. But as long as the sort of probability used is not subjective probability, this poses no major philosophical problems. (For a consideration of some of the technical problems, see Hurley 1989 and Gardenfors and Sahlin 1988.)

One last complication with means-ends analysis is that something is not the best means to an end if it is itself impossible to achieve. Suppose that the answer to the question, 'How should I get to the shop?' is, 'By jumping over my house'. The whole point of means-ends analysis is to break down a task into a *possible* way of doing it.

All means-ends analysis whether for a company profit-maximizing or for a selfish agent's pleasure maximizing or for a moral consequentialist agent's good maximizing depends on a repertoire of possible activity. There are certain things which can be controlled by a company, and the practically rational courses of activity of the company must be built up out of these.

How do we decide what should be in the basic repertoire? *We* don't. The practical rationality mechanism determining the activity of some agent works according to a theoretical method embodying practical rationality that is relative to some repertoire of basic activity. What is or is not basic activity is ultimately determined by how that mechanism works. There might be an indefinite number of theoretical methods embodying practical rationality that only differ in terms of what basic repertoire they use. There is no theoretical reason to prefer any one of these theoretical methods. The one that is relevant is the one that represents how we do actually work.

The same goes for the constraints on the kind of objective facts that can feed into a method of practical rationality. Some methods may take in any facts, whereas some may take in only locally apparent facts. Again, the one that will turn out to be relevant is the one that represents how the agent actually works.

On my account, practical rationality is relative to a theoretical method but not relative to a set of fundamental desires. Knowing a theoretical method is like understanding a language. It involves being able to apply a structure of concepts to derive an indefinite number of

different descriptions including those applying to novel situations. Moral argument can be seen as operating either within a theoretical method or across theoretical methods.

This does not stipulate against moral absolutism. Indeed, I think that this provides a basis from which to consider such questions. There are several ways that one might argue for absolutism from this account. The first is to argue that given a reasonably powerful set of concepts only one coherent theoretical method is possible. (Or only one class of intertranslatable theoretical methods is possible.) A second way is to argue that the moral 'should' is stronger than the 'should' of any particular theoretical method. Possibly there are some basic principles for improving theoretical methods (either these principles are encapsulated somehow in the theoretical methods themselves or they involve meta-considerations like descriptive inclusiveness etc.) Absolutism would require that there was an ideal result of theoretical method development, and that what you should do morally is what this ideal theoretical method would say you should do. To argue for this one would have to show that there was only one path a theoretical method might take or that theoretical methods were bound to converge, and that there is an end point to moral development (i.e. development of the theoretical method of practical rationality).

This is all just supposed to be intuitive support for regarding practical rationality as relative to a theoretical method. Although the position is clearly incompatible with any view that grounds moral reasons in desire, it is in itself neutral with respect to questions of

moral realism and moral absolutism. Indeed, although my account of practical rationality as objective suggests that moral rationality may be part of practical rationality, one can quite consistently hold that moral rationality involves something quite different in addition to means-ends analysis.

It may be objected that requiring that practical rationality embodies means-ends analysis at all commits me to the position that all rational agents are consequentialists. But I do not want to be so committed. There is something involved in human practical rationality which seems to go well beyond simple means-ends analysis. This is inevitable if there are competing top-level goals or perhaps no fixed top-level goals at all. By working with a notion of objective practical rationality as *just* embodying means-ends analysis I have been trying to make my case with a minimal notion of objective practical rationality. But, whatever advances one cares to make on this notion should be easily incorporated into the account as long as they do not require that practical rationality essentially takes into account beliefs and intentions.

In summary, there seems to be no theoretical objection to a theoretical method embodying objective practical rationality. We can provide simple unproblematic examples like a theoretical method which describes a company's activity leading to profit maximization as practically rational. It is an interesting and important question whether or not fixed top-level goals can be eliminated from such theoretical methods by making them more sophisticated. But that is not

the issue on which my account stands or falls. The crucial issue is whether it makes sense to account for the way people do operate in terms of any such theoretical method embodying objective practical rationality.

5.3 Objective Practical Rationality Mechanisms

The intuition at work against my account is that a theoretical method involving rationality considerations which are *external* to a person cannot represent how that person works. Only a theoretical method which is sensitive to the internal states of a person can have any chance of representing how they work. Only such a theoretical method could represent how it really is with them. According to this intuition, it is irrelevant whether a theoretical method embodying objective practical rationality is possible since such a theoretical method cannot have any role in explaining how people *actually* behave.

I hope that I have already gone some way towards diffusing the force of this intuition with my account of causal explanation. For example, one of Hume's arguments for subjective rationality just does not get a foothold given this account. That is the argument that any explanation of action must make reference to the motive force; and that motive force can only be a desire (passion). Otherwise it is like trying to explain how a car works without referring to the petrol.

On my account of causal explanation, a good reason must have a logical place in a theoretical method that represents how the underlying mechanism works. The theoretical method that represents the way the mechanism works should not be sensitive to considerations of whether or not the mechanism is present. So it need not take into account the motive force of the mechanism. Of course, the motive force must be there supporting the explanation. But this fact need not be incorporated in the theoretical method itself, but should rather be incorporated in the fact that there is a mechanism present whose working is represented by the theoretical method. (The same applies against the similar argument that beliefs need to be involved in an explanation of action.)

Moreover, invoking desires cannot explain how our bodies are wheeled into action unless there is a mechanism underlying Humean practical rationality that transforms desires into actions. It must always be the mechanism underlying the explanation that provides the motive force.

On my account in chapter 6, desires are ascribed *in virtue* of the presence of a mechanism that determines activity in line with practical rationality. So, given my account, invoking a desire does after all amount to invoking the motive force. This is why the Humean thought about the need for invoking desires sounds attractive. But it is important to see that desires fulfil this motivating role in themselves only in virtue of there being a way of explaining activity which does not rely on desires as the motivating forces.

So, I am recommending the move of taking desires out of the teleological explanation of activity and only bringing them back into the picture in virtue of there being such teleological explanations. And of course I am not denying that we can then invoke them in explaining behaviour once we have brought them back in. In chapter 6, I argue that there are rather uninteresting and non-teleological explanations of activity in terms of beliefs and desires. My claims are that we *can* explain activity without reference to desires, and that this sort of explanation is fundamental.

Now consider the objection that there is no real difference between us and the 'smart' computer systems considered in 4.4. If it is too difficult to work out what is the best thing to do, we will not do what is best either. Suppose that I come to a junction in the road and the place I am headed for is signposted to the right. Unknown to me, and practically unknowable to me, the sign has been swivelled round and I should have turned left.

In this case it seems that I am not really adaptable to what is objectively rational to do. This can be generalized to the conclusion that nothing can operate according to objective practical rationality, because no system can have an unlimited capacity to discover what is objectively best.

It might be thought that one can deal with this objection by some restriction of the operational conditions of an objective practical rationality mechanism. Perhaps the operational conditions of a mechanism that works this way should include the conditions that the

information concerning what is practically rational is available to the mechanism, I.e. the situation is such that someone should be able to find out what is practically rational. So the mechanism is not wholly internal; it is externally directed. It has external operational conditions, in particular that the information leading to finding out what is practically rational is available.

But the argument of 4.4 against this move can be made here. For, this restriction on the operational conditions of the mechanism rules out the possibility of some parts of the characteristic structure of events of the mechanism ever occurring. This means in turn that the description of the working of the mechanism that includes these characteristic results is inaccurate.

So tinkering with the operational conditions of a practical rationality mechanism will not get round this objection. The theoretical method representing how such a mechanism works needs to be tinkered with instead. This alteration could be done in one of two ways. We could say that the mechanism produces objectively practically rational activity but only when information concerning what is objectively practically rational is available to the mechanism. Or, we could alter the notion of practical rationality itself, and say that what is practically rational depends on what information is available.

According to the first sort of alteration, it is rational to go right even though the signpost points left; but the practical rationality mechanism does not necessarily produce rational results when the necessary information is not available. According to the

second sort of alteration, it is rational to go left even though the goal is to the right, because that is what the available information suggests. This second way has the advantage of explaining why I do in fact turn left.

The worry now is that such necessary tinkering means that the theoretical method no longer embodies *objective practical rationality*. The objection sees two possibilities corresponding with the two sorts of alteration of the theoretical method. Either we are left with the description of a mechanism as characteristically resulting in activity that is objectively practically rational, but only when it *seems* to the mechanism that such activity is objectively practically rational. Or we are left with the description of a mechanism as characteristically resulting in whatever activity *seems* to the mechanism to be objectively practically rational. In either case, subjective considerations have become central to the notion of practical rationality. Indeed, in the second case there is no vestige of objectivity left.

This objection is based on the assumption that the only way of understanding the notion of information being available to a system is subjectively - i.e. in terms of the system being aware of that information. It is fairly easy to see that this assumption is false. Suppose that there is a correct signpost in front of me showing me where to go, but I neglect to look at it. In this case the relevant information is available but I am not aware of it. The fact that the information is available is not subjective.

It might be argued that whether or not information is available to me depends on whether I have the perceptual or reasoning machinery to discern it. But this is exactly parallel to the consideration that what activity is possible for me depends on what behavioural machinery I have. In both cases there are as many different theoretical methods as there are assumptions about what basic repertoire the practical rationality is relative to. If one of them represents how I work, then *my* basic repertoire is thereby determined.

It is not my task here to analyse the notion of information being available. The only result I require is that a theoretical method may involve a way of telling whether the information that X is practically rational is available to a system without requiring any input from the subjective state of the system. I trust that I have this result.

Even if the notion of information being available turns out to be a sort of secondary quality, this does not make it subjective. It might turn out that the notion should be thought of in terms of something like obviousness. Whether and to what degree some piece of information is obvious may depend on whether it would seem obvious to a normal observer. But it would not follow that a situation has the property of obviousness in virtue of the subjective state of the *agent in question*.

5.4 Partial Practical Rationality Mechanisms

Suppose I misremember how to wire up a plug properly, and attach the blue wire to the live terminal. According to Humean practical rationality this activity is rationally explained in terms of the belief that the blue wire is connected to the live terminal when the plug is wired properly and the desire to wire up the plug properly. So, according to Humean practical rationality, the activity is the result of a practical rationality mechanism working properly. This is a desirable conclusion because the activity is after all purposeful.

On the other hand, there is a failure of rationality here also. This might be accounted for by a Humean as a failure in the mechanism that results in rational beliefs given experience. But the Humean must stop applying rationality at this point. Experience is not rationally assessible for a Humean. It is the bottom line. This leads to the familiar problem of scepticism. The problem is that reasons, on the Humean model, cannot justify beliefs about the world; they cannot take an individual beyond the purely subjective realm.

For an externalist, perception itself should be regarded as an active process of gathering information. In particular, perception (along with reasoning etc) should be seen as part of the larger practical rationality process. And here the bottom line for practical rationality is the way the world is; it is not a subjective state of the individual. If someone is walking down the street, it is rational for them to find out if there are any obstacles in front of them. (A

Humean might argue that this is based on the *belief* that finding out whether there are any obstacles is necessary to walking down the street. But is this belief itself rationally assessible?)

So much for the *irrationality* of my activity as I wire up the plug wrongly. Now consider its *rationality*. Humeans seem to have this under control for they base the rationality of the activity on subjective states. But the externalist cannot do this. It seems that the objective practical rationality mechanism is not working and the purposefulness of the activity cannot be accounted for. In this example we have activity which is purposeful and objectively irrational at the same time. Note, this activity will be irrational however weak a notion of objective practical rationality is being employed. So my task seems to be to show how objective practical rationality mechanisms can be working (to account for purposefulness) and not working at the same time (to account for irrationality).

When I misremember how to wire up a plug something is wrong with my practical rationality mechanism. If it were put right, then my activity would become rational. The failure exhibits itself in the activity, but what was wrong with the mechanism was present (or absent) earlier. Despite the failure of the mechanism in this respect it works properly before and after this lack and indeed also during it with other aspects of activity. So there is no problem in identifying that such a mechanism is operating even when it partially fails.

Consider the mechanism responsible for courting behaviour in grebes. Suppose that something goes wrong and the activities of a pair

of birds gets out of synchronization. The mechanism, which when it works properly, results in the pair building a nest and mating, has some other more disappointing result - failure and sexual frustration. The failure can be attributed to a local lack in the operational conditions at some stage of the process. But this does not stop the mechanism operating - it just means that the changeable part of the underlying nature of the mechanism is no longer the way it should be for the process to continue as it should.

Similarly with purposeful activity; if something goes wrong with the perceptual stage of the working of a practical rationality mechanism, then all the other stages go out of synchronization. The practical rationality mechanism is operating throughout, but the result of this operation is not what it should have been if nothing had gone wrong. The characteristic results of the practical rationality mechanism do not materialize even though it is operating.

This may suggest the thought that the purposefulness of objective irrational behaviour is to be accounted for in terms of the fact that that behaviour is the result of an objective practical rationality mechanism working, but not working properly. (See the end of 2.5 for an account of what it is for a mechanism to be working improperly.) But I think that this thought is a mistake, because it separates purposefulness from teleological explanation.

Irrational activity resulting from an objective practical rationality mechanism not working properly cannot be explained in terms of the way that mechanism works. The description that I put the

blue wire to the live terminal is not determined by this theoretical method embodying practical rationality. The thought was that this activity is purposeful in virtue of being the result of a practical rationality mechanism whether it works properly or not. Yet the activity cannot be explained in terms of how this mechanism works. The reason for my strange behaviour is not that such behaviour is the best thing to achieve according to the theoretical method. For it is not. There is no teleological explanation of my behaviour in terms of this theoretical method embodying objective practical rationality.

If purposefulness did not require teleological *explanation*, but just causation by a teleological mechanism, then the problem of deviant causal chains would come back. If an ill-functioning objective practical rationality mechanism by luck resulted in the objectively rational activity, we would not want to describe that activity as purposeful. There was purpose behind it, but that is a different matter. To count as purposeful, the activity must be the result of a teleological *process*. This means it must be explainable in terms of practical rationality.

This seems to leave the door open for the internalists. They can explain my irrational behaviour in exactly the way they explain my rational behaviour, for both are in fact rational on their conception. And they can maintain the attractive position that the purposefulness of behaviour is a characteristic of the way that behaviour is explained.

However, on my account, there *is* a way of explaining irrational behaviour, even if it is not with respect to complete objective practical rationality as such. When the objective practical rationality mechanism is working but not working properly, we can identify other mechanisms that *are* working properly. These mechanisms embody the failure in practical rationality as well as embodying objective practical rationality in other respects. The mechanisms only operate as long as the flaw in the operation of the objective practical rationality mechanism lasts. I shall call these 'partial practical rationality mechanisms' or 'partial teleological mechanisms'.

The way they work may be represented by the following kind of theoretical method. Start with one's basic theoretical method embodying objective practical rationality. In a situation where there is no information available about the truth of B_1, B_2, ... or about the rationality of D_1, D_2, ..., then describe as rational what would be practically rational according to the basic theoretical method if B_1, B_2, ... were true and D_1, D_2, ... were practically rational. Where B_1, B_2, ... conflict with the way the world is and D_1, D_2, ... conflict with what is objectively practically rational according to the basic theoretical method, then B_1, B_2, ..., D_1, D_2, ... should be regarded as dominant. (This specification of a theoretical method embodying partial objective practical rationality will be refined in chapter 6.)

In the wiring example, B_1 could be that the blue wire is supposed to be connected to the live terminal. Given it is practically rational to wire up the plug properly according to the theoretical method, then

the theoretical method would determine the description that I will attach the blue wire to the live terminal. If this theoretical method represents the working of the mechanism that does result in this behaviour, then we have an explanation of my behaviour. My hand moves in a certain way because that is part of the way to attach the blue wire to the live terminal. I attach the blue wire to the live terminal because the plug is to be wired up properly. What we cannot say is that I attach the blue wire to the live terminal because that is the way to wire up the plug properly.

Given that one can explain my behaviour with reference to this partial practical rationality mechanism, what is the point of including the complete practical rationality mechanism in the account? One answer is that the partial practical rationality mechanism is parasitic on the complete practical rationality mechanism. The partial mechanism includes as part of its essential nature the failings in the complete mechanism. When the operational conditions of the complete practical rationality mechanism are satisfied the partial practical rationality mechanism will cease to be. When I find out how to wire up a plug properly the mechanism that works as if blue should be attached to live will disappear. The essential nature will be no more. Another partial practical rationality mechanism working in a slightly different way will take its place. The sequence of partial practical rationality mechanisms owes its existence to the complete practical rationality mechanism.

Note that the mechanism underlying the 'smart' computer system considered in 4.4 is not a partial practical rationality mechanism. The theoretical method representing how it works has nothing but fixed points, B_1, B_2, etc. So its activity is made 'practically rational' because of the fixed points alone without any adaptability to the way the world really is.

The image that one should have of these various practical rationality mechanisms is the following. The complete practical rationality mechanism is the important one. It may or may not change gradually over time depending on your moral philosophy. It is constantly working, though quite often, because of certain interferences or local failings in the operational conditions, it doesn't work properly. Entirely dependent on this mechanism we can identify a shifting series of partial practical rationality mechanisms which operate properly when the complete practical rationality mechanism goes wrong. These are opportunistic mechanisms which owe their existence to the failings of the complete mechanism. When the complete mechanism works properly, which it must do when its operational conditions are restored, the opportunistic partial practical rationality mechanism disappears, to be replaced immediately by another one a bit closer to the complete practical rationality mechanism.

These partial practical rationality mechanisms do the work that Humean practical rationality is supposed to do. They explain objectively irrational activity. (I will not capitalize on it until chapter 6, but I have just constucted the basis for my theories of

belief and intention. A subject believes B if and only if their activity is determined by a partial practical rationality mechanism which works as if B is the case, and intends to achieve D if and only if their activity is determined by a partial practical rationality mechanism which works as if D is practically rational.)

Chapter 6

Philosophy of Mind

6.1 Purpose and Agency

In the standard causal theory of action considered in chapter 1, intentions are regarded as basic, and teleological explanation is just explanation in terms of intentions. There is no doubt that the concept of intention is bound up with both the concept of action and the concept of teleological explanation. My suggestion in chapter 1 was that the causal account had the direction of analysis the wrong way round; that teleological explanation was the more basic notion, and that the notion of intention is to be understood in terms of that. I have constructed the account of teleological explanation without reference to inner mental states. Now, to complete my reversal of the standard causal account, I should show how we can understand intentions in terms of such a notion of teleological explanation.

In the process of turning the standard account upside down I pass through many of the main areas in the philosophy of mind. As I go, I sketch how my account may be applied to these areas. To claim to do anything serious in this respect in a single chapter is little more than a joke. The aim is really twofold. One task is to show that my account does not fall at the first fence with any of the problems in

the philosophy of mind. The other task is to provide a glimpse of the potential of this account and to provide a rough programme for future research. If the account is correct, then it can do some wonderful things. If not, then there is no point in spending more than a single chapter in showing what it could do if it were correct. So, I largely ignore the philosophical literature on these subjects and make little effort to criticize any alternative positions.

The first mentalistic notion to emerge from my account is that of purpose. The purpose, as I rather austerely interpret the term, is just the goal of a teleological explanation. If some activity happens because it is a means to achieving some end, then that end is the purpose of the activity. I think that there is a reasonably clearly defined sense of the word "purpose" which is being explained here. According to this sense, for some activity to have some end as its purpose it must actually be a means to the achievment of that end. To make the austerity of this sense more explicit, it might be better to say that this is the notion of some activity *serving* a purpose.

We also talk of a person *having* some purpose in doing something. In this sense, the activity need not actually be a means to the end. Here, the idea of a *partial* teleological mechanism is important. The activity has some end as a purpose in this sense if the activity happens because it would be a means to achieving the end if the world were different in some respect. My purpose in walking to the bank last Sunday was to cash a cheque even though my activity certainly did not serve that purpose. Had the bank been open, my activity would have

been a means to that end. The mechanism determining my activity was sensitive to what comes out as practically rational according to the method of partial practical rationality embodying the false proposition that the bank was open.

In this sense, the purpose is the intention. This will be dealt with more completely in 6.3.

Here is my simple theory of agency. Any system with purpose is an agent. Purpose is here to be understood in the austere sense. (It is difficult to disentangle the two senses as requirements of agency since any system with purpose in this sense also has purpose in the sense of having intentions.) To put the theory of agency bluntly, anything whose activity is run by a teleological mechanism is an agent. People are essentially teleological machines.

It is quite often thought that, in addition to having purpose, an agent is essentially something that generates its own purpose. (C.f. Taylor, C. 1985 pp15-44 and Frankfurt 1971) One pressure towards this thought is that computers have derived purpose and are not agents. However, I hope I have shown that such computers may have functions, but do not have purposes, derived or otherwise. Once the pressure of such counterexamples to the simple theory are removed, the requirement that agents must derive their own purpose seems unmotivated and unduly strong. Agents might have some top-level goal for psychological reasons that are not in their power to alter. In this case, they do not determine their purpose, but they are agents for all that.

Note the powerful idea underlying theories of agent causation. (C.f. Chisholm 1966, Taylor, R. 1966) For activity to be part of an action, it must derive from the agent as source. This simply falls out of my theory of agency. According to the theory, purposeful activity is activity resulting from a teleological mechanism. If an agent is essentially run by a teleological mechanism, then activity is purposeful just when it results properly from the essential nature of an agent.

Is this sort of causation essentially different from event causation? The fact that this is mechanism-causation rather than event-causation is a trivial respect in which it is. But there is nothing special about mechanism-causation that distinguishes the causation of purposeful activity from the causation of unpurposeful activity. For example, one can explain the fact that water is sucked into the roots of a plant as a result of the transpiration mechanism of the plant rather than as the effect of some cause.

What is special about agent causation is that it is teleological. One cannot account for the purposefulness of activity by looking at it as the result of a non-teleological process taking mental states as inputs and issuing activity. Only by looking at activity as the result of a teleological process, can one see the central role of the agent - the teleological machine.

Assuming that a person is an agent, then, according to my account, a person is essentially something run by a teleological mechanism. This seems to capture several influential thoughts concerning the nature of a person. First of all, there is the ancient thought that a

person is essentially a rational being. This is exactly what I am saying. I am not however saying that a person is a rational *animal*. Unlike Wiggins for example (1980 ch 6), I think it is as much a mistake to think that we are animals as it is to think we are bodies. When you look in the mirror you see yourself, you also see a body, and you also see an animal, all occupying the same space. But this does not mean that these are one and the same entity. I suggest that if your brain was damaged so that you lost all powers of rational behaviour, you would be no more, but the animal might be living and breathing.

The account is surprisingly close to Locke's.

> *'[a person is] a thinking intelligent being that has reason and reflection and can consider itself as itself, the same thinking thing in different times and places; which it does only by that consciousness which is inseperable from thinking and, as it seems to me, essential to it.'*
> (Locke Essay II 27.9)

What Locke's account includes, in addition to the thought that a person is essentially a rational being, is the essential role of consciousness. Now, it may be that consciousness should be understood in terms of an agent's teleological mechanisms. This is not something I intend to show here. But whether or not consciousness is to be understood in this way, it is not the case that consciousness is strictly essential to being an agent (i.e. having a teleological mechanism determining one's activity). The potential for consciousness may be essential, but consciousness itself is not. You do not have to

be conscious to be a person; you might be asleep. So, even if the potential for consciousness is a result of what is essential to a person, it does not follow that consciousness itself is essential.

But is my suggestion, that a person be essentially a rational being - i.e. a teleological machine - itself too strong? It would mean that probably very young babies and certainly unborn babies were not people, despite being human beings. The same would go for adult human beings with massive brain damage who had lost all capacity for agency.

I can find no reason to doubt these conclusions. Nor do I have any real objections to counting Martian agents or animal agents as people. But I am certainly not committed by my account to this last conclusion. For, although I must accept that people are agents, I can allow that they are special kinds of agents - human agents for instance.

Regarding people as essentially embodying teleological mechanisms provides a natural way to explain the individuation of people. If a single body was run by two entirely separate (metaphysically independent) teleological mechanisms, then two people could be identified. The re-identification of a person over time can also be understood in terms of the underlying teleological mechanism. At any one time a number of teleological processes will be happening simultaneously as a result of the working of a teleological mechanism. These processes overlap and interact. Even when the teleological mechanism changes, so long as it is gradual, many of the current processes will continue and thus maintain continuity. So, although a

person may be operating by a totally different teleological mechanism after some years, if the mechanism has been changing continuously, then the person can be identified as the same as before. A single path through time can be identified, not simply as the path of a body, but as an overlapping sequence of goal-directed processes with a continuously developing underlying nature.

Amnesia does not fatally interupt the process of an agent's course through life. For enough of the same mechanism is still working. Indeed, many of the goal-directed processes continue unchecked. However a very radical and sudden personality change might give one reason to say that a new person had come into being. The personality of an agent is roughly the way the agent's teleological mechanism works; so when one personality disappears suddenly to be replaced by another different one, the same can be said of the person.

6.2 Establishing the facts

Traditional accounts in the philosophy of mind regard images, thoughts and intentions as basic entities, and build up accounts of perception, reasoning and action in terms of these. My general criticism of such accounts (apart from the fact that none of them actually works) is that they leave the unanalysed mental entities (images, thoughts, etc) lying undigested at the bottom. The real questions become: what are images, beliefs, intentions, etc; how should

they be attributed; in virtue of what do they have the content they do? None of these questions can be answered by the traditional mode of analysis, because, by regarding these entities as basic, the analysis just is ducking these questions.

The traditional way of looking at the whole action process is as the product of three parts; the formation of images and basic beliefs (perception); the formation of less basic beliefs, plans and intentions (reasoning); the formation of behaviour (action). One starts with the basic mental building blocks and puts them together to construct a story of how people operate. But I am argung that we are essentially teleological machines. If this is so, then the traditional account must be misguided. For, it is clear that anything that can perceive, reason and act is an agent. If these processes can be understood as possible independently of being part of a teleological process, then the teleological process is not an essential part of what it is to be an agent. So, I must urge that these processes can only be understood for what they are if they are seen as essentially being parts of potential teleological processes.

To that end, I start with a notion more immediately understood in terms of potential teleological processes - the notion of establishing a fact. The traditional accounts would analyse the notion of establishing that p in terms of perception and reasoning. I go the other way round. Establishing that p is a process that represents a potential advance in teleological processes. The advance is characterized in the following way. Before you have established that p

you may need to be presented with the fact that p if you are to behave rationally. Afterwards you will not need to be. This is just what it is to establish that p. A process, X, is said to be the process of establishing that p if and only if there might be a teleological process that was incomplete because the fact that p was not apparent to the agent, and X would make it complete.

Suppose that I am staring gloomily out of the window at the rain. According to my account, I have been establishing that it is raining even though I do nothing that uses the fact. I have been establishing that it is raining in virtue of the fact that *if* the rationality of some action depended on the fact that it was raining (for example, if I had the goal of catching pneumonia by sleeping in the rain), but I failed to do it because the information was not available, then the process I have been undergoing would fill the gap and enable my teleological mechanism to work properly after all. In virtue of staring out of the window I can act rationally with respect to the fact that it is raining.

So establishing a fact can be done independently of any particular teleological process. There need be no *manifested* advance in one's rational achievement of goals. But there must be an advance that is at least *manifestable*. My claim is that one could not correctly attribute the establishing of a fact (and hence perception or reasoning) to some agent unless one could attribute a change in the agent's partial teleological mechanisms. And such changes will always be at least manifestable.

I think that perception and reasoning should be regarded as special ways of establishing the facts. Perception essentially involves whatever is being established being presented to the sense organs. Reasoning does not. Of course, this idea is no more than a rough plan for real accounts of perception and reasoning. My aim is just to suggest the possibility of not doing an account in the traditional way which leaves undigested mental residue.

In favour of an account such as mine is the thought that the content of perceptions can only be individuated by some sort of discriminating capacity, and such a discriminating capacity is only manifestable in action. Only through dispositions to behave in certain ways can it be determined whether I am perceiving a chair say rather than only perceiving pieces of wood. The content cannot be determined by the inputs since information from a chair may be coming in without the chair being perceived. It cannot be regarded as intrinsic to the mental representation, since this is nonsense. So, it must be determined by what the agent is liable to do with the information.

This sort of thought is expressed by Evans for example about concept possession in general.

> *'the concept of knowing what it is to be true that p ... is one of a* capacity, *and the proof of its being possessed at a given time must surely reside in facts about what the subject can or cannot do at that time.'*
> (Evans 1982 p116)

Associated with this thought is the psychological fact that you cannot see some aspect of a thing if you cannot imagine doing anything

that depends on that aspect. Witness the way children start to see the world. They certainly do not see everything in a scene at once. They just see what they need to see or might need to see. One learns to see some aspect (a colour for example) by learning how to discriminate that aspect in one's goal-directed behaviour.

It is important to see that this is not a purely cognitive account of perception. Perception is not being analysed in terms of judgements and beliefs, but in terms of something more basic - behavioural mechanisms. So, for example, Peacocke's arguments (Peacocke 1983 ch 1) for there being a non-representational content in perception do not really challenge the account. Peacocke argues that there are features of perception which are not included in the way the world is represented to the subject. Such features might be the special feeling associated with binocular rather than monocular vision or the sense of how large an object is relative to the subject's visual field.

The perceptual difference between looking at a scene with one eye and looking at it with two may be accounted for as a difference in the change made to the subject's teleological mechanisms. There is a subtle difference in the degree of confidence the subject would have in making depth judgements during and after the two experiences. Similarly, the feeling that a closer tree is in some sense larger (in one's visual field) than the same tree further away involves a characteristic change in one's teleological mechanisms. If, for some bizarre reason, one needed to obscure the sight of the tree, one could do it without any

more information being available. So, size relative to visual field can be part of perceptual content on a behaviourist account.

Peacocke intends these examples not just to show that there is non-representational content in perception, but more to show that there is content which is not to be accounted for in terms of cognitive processes. But the examples evidently fail to show this. The same goes for other standard criticisms of cognitive accounts of perception.

For example, it is possible to perceive some unchanging scene and then a moment later still perceive it, but all the facts to be established had already been established in the first moment. So what sense can it make on my behaviourist account to say that the scene is still being perceived? One answer is that the fact that the scene is *still* present is constantly being established. But what if I knew it was still present from other sources? This consideration is irrelevant so long as it is always possible to imagine a situation in which I did *not* know that it was still present from other sources. Considering the difference my current process would make in such a situation is all that is required to determine whether this current process is a process of establishing that the scene is still present. A change in one's teleological mechanisms can be identified even when the change is not actually reflected in one's beliefs.

The same goes for cases where, because of false beliefs, I fail to act on information received. We can still say that perception is going on as long as the process *would* fill the appropriate gap in some

potential teleological process that only lacks the information being apparent.

6.3 Intentions and Beliefs

In the normal working of a complete teleological mechanism there comes a stage at which no more perception or reasoning concerning the truth of a proposition p is required for the agent to be able to act rationally in situations where the truth of p is relevant. This does not mean that the agent will be able to act straightaway in these situations without any more facts being established (perception and reasoning going on). For, almost invariably, information concerning other matters will need to be acquired. But at this stage, no more advances in the process which can be attributed to the specific ability of the agent to act on the truth of p are required. At this point we may say that the agent believes that p.

Suppose now that the complete teleological mechanism goes wrong. p is not true, but the agent is presented with an illusion that p is true. When the complete teleological mechanism is working it gets to the bottom of the illusion. But, for one reason or another, it fails in this respect. The result is that the agent's way of behaving is very much like it would have been if p had been true and the agent had correctly perceived that. In fact, in those situations where there is no more perception or reasoning concerning the truth of p, the agent's

way of behaving is just the same. Here also we say that the agent believes that p.

Exactly parallel considerations determine how we attribute intentions to an agent. If the agent has reached a stage where no more perception or reasoning concerning the practical rationality of q is required for the agent to be able to act rationally in situations where the practical rationality is relevant, then we may say that the agent intends to achieve q. Also, if the agent's way of behaving is just the same as this where no more perception or reasoning concerning q happens, even if the behaviour is irrational since q is not practically rational, we may still say that the agent intends to achieve q.

This can all be characterized more precisely in terms of partial teleological mechanisms. Remember that these mechanisms are not metaphysically independent of the complete teleological mechanism. Their ways of operating are characterized in a slightly different way with the result that even when something goes wrong with the working of the complete teleological mechanism in the perception or reasoning part of the process, some partial teleological mechanism will still turn out to be working properly. This is basically because that part of the perception or reasoning that goes wrong in the complete mechanism is excluded from being part of the partial process.

The way a partial teleological mechanism works is as follows. In situations where no more perception or reasoning concerning the truths of B_1, B_2, ... or practical rationality of D_1, D_2, ... happens, then what will happen is what would be practically rational according to the

complete theoretical method were B_1, B_2, ... true and D_1, D_2, ... practically rational. Describe B_1, B_2, ... and D_1, D_2, ... as fixed points of the partial teleological mechanism. Then an agent believes B or intends to achieve D if and only if they are run by a teleological mechanism with B as a fixed point or D as a fixed point.

Note that a partial mechanism like this can fail like any other mechanism due to failure in operational conditions. For example, there may be a failure of attention to certain information. But one lack that cannot account for its failure is a lack of perception or reasoning concerning the truth of B_1, B_2, ... or the practical rationality of D_1, D_2, ... The point established in section 4.4 is relevant here. The operational requirements of a mechanism cannot be such as to make redundant part of the theoretical method, or else it makes no sense to describe the mechanism as operating according to that very theoretical method. It is part of the theoretical method of this partial teleological mechanism that no more establishing of the facts concerning B_1, B_2, ... or D_1, D_2, ... occurs. So a lack of such perception or reasoning cannot be taken as an excuse for the mechanism not working.

The way I have set up the definition, there is no need for all of an agent's beliefs and intentions to be incorporated in a particular partial teleological mechanism. This is not a problem, because if an agent is run by a partial teleological mechanism with X as a fixed point and at the same time run by a partial teleological mechanism with Y as a fixed point, they will also be run by a partial teleological

mechanism with both X and Y as fixed points. The proof of this goes along the following lines. In situations where X and Y would combine to make something else rational, the agent acts on the basis of the combination. This is because the agent's complete teleological mechanism is such that it would be working properly if X were true or practically rational in situations where this was relevant, and is such that it would be working as if Y were true or practically rational in situations where this was relevant. So in situations where both are relevant, what happens is what would happen if they were both true or practically rational and the mechanism was working properly.

An extension of this argument suggests that at any one time there is a single partial teleological mechanism operating an agent that incorporates as fixed points all the agent's beliefs and intentions. This can be thought of as a maximal partial teleological mechanism. It is crucial to realize that this maximal partial teleological mechanism, although it operates in situations where there is no more perception or reasoning going on concerning any of these fixed points, nevertheless still essentially involves perception and reasoning when it does operate. There will always be an infinite amount of things which one does not have any beliefs or intentions about. In the simplest action like turning on a light switch, finding out about some of these things will be necessary if the goal is to be achieved. For example, perception concerning whether or not one's finger is in the right place is absolutely essential to this action.

So I maintain vigorous resistance to the Humean thought that also underlies most functionalist accounts that behaviour can be rationally explained by reference to beliefs and intentions alone. The thought is that the system of beliefs and intentions results from the process of perception and issues in behaviour through the process of practical reason. This system is seen as an isolatable segment of the whole process. So no non-subjective reasons need be considered in the explanation of action. What is clearly wrong with this is that even a maximal partial teleological mechanism needs help from the world in the form of objective reasons to account for even the simplest acts.

It might seem neater to analyse beliefs and intentions in terms of a single maximal partial teleological mechanism. But I can see no good reason for this and some against. How do I establish whether I believe that p? One way is to consider what I would say if it were rational for me to speak the truth on this matter and assuming I do no more perceiving or reasoning concerning the truth of p. But this is not considering how my maximal partial teleological mechanism works. For, it is a fixed point of that maximal mechanism that it is not rational for me now to speak the truth concerning p. (I am sitting in a library; I don't intend to open my mouth.) What I do is imagine that some other partial teleological mechanism that *has* got the practical rationality of speaking the truth now is operating and consider whether p is a fixed point of it. Another reason not to put too much stress on the notion of a maximal partial teleological mechanism is that every second that my eyes are open a new maximal partial teleological

mechanism takes the place of the old. In the process of working the mechanism ceases to be maximal.

There is a principle underlying my account that it makes no sense to attribute any mental state to oneself or others if that state is not manifestable somehow. I can find no example of a correct attribution of belief or intention that puts pressure on this principle. It might be thought that a problem is posed for my account if someone who has no belief concerning the truth of B decides nevertheless to act just as if B were true anyway. But, according to this hypothesis, the person would have to act as if B were true even in situations where a lot depended on getting it right. Now, to make sense of this situation at all, there must be some reason for the agent behaving in this way. For, if not, then there could be no motivation for denying that the agent believed that B. But if there is such a reason, then we can devise a situation in which that reason fails to hold and the true nature of the agent's behaviour is manifested.

A related question is that of *dormant* beliefs and intentions. Can my account distinguish them from merely *potential* beliefs and intentions? By a dormant belief I mean a belief that is not currently motivating any behaviour at all. For example, a few minutes ago it was true that I believed that grass was green, but having that belief was not manifested in my behaviour. I can be said to have had that belief because it had the potential to be manifested in my behaviour had the situation been appropriate to that. On the other hand, that there is a solitary dandelion in the middle of the lawn also had the potential to

be manifested in my behaviour. But I did not believe it. The difference is that before the fact that there is a solitary dandelion in the middle of the lawn can be manifested in my behaviour I have to look out at the lawn, whereas no such perception is required for the fact that the grass is green to be manifested. Some perception is required of course even in this case. But the perception concerns things like the position of the piece of paper in front of me and not the greenness of the grass.

The distinction simply falls out of my theory. One of the things that distinguishes the way a partial teleological mechanism works from the way a complete teleological mechanism works is that the partial teleological mechanism produces its results without any more information being acquired about the fact in question (one of the fixed points of the mechanism).

Now consider potential beliefs which require a certain amount of reasoning in order to be formed and then manifested in behaviour. For example, in a game of chess I did not believe when I started thinking about the move that I should sacrifice my queen, but after a few minutes of thinking I did believe this. Suppose a situation had arisen when I first started thinking about the move in which I had no more time to think before acting. In other words any more establishing of the facts about the best move was ruled out. The result would not have been in accordance with a partial teleological theoretical method working as if sacrificing the queen were rational. (Note that even if I hit on the move by luck I would not put money on it.) There is no

failure in operating conditions that can account for this failure. So I simply do not at that point operate according to such a theoretical method.

So it looks as if my theory accounts for the distinction between actually having beliefs that are not currently being manifested and only potentially having beliefs. Can it work as well for the same distinction with intentions? Cases where an agent only potentially has an intention (e.g. were they to perceive some relevant facts or do some relevant reasoning) are ruled out as real cases of intention on my account in just the same way as potential beliefs are. It might be thought that the notion of a dormant intention was more problematic than the notion of a dormant belief however.

Suppose that I form the intention to have a massive birthday party on my 80th birthday. On my account, this must involve having a current mechanism capable of producing results *now* that would be rational were it rational to have such a party. But there is an apparent objection that the having of this intention may have no way of being manifested in current behaviour. According to the thinking behind this objection, one has the intention in virtue of having certain thoughts - mental reminders to be picked up in the future - and these need only be tied to behaviour in the distant future in order to constitute the having of a real intention.

However, suppose the situation arises in which it becomes apparent that my memory will not be sufficient to the task. If I really have this intention to have an 80th birthday party, I will act on it now by

making an actual note of the intention and somehow making it difficult for myself to get out of the committment when the time comes. If I would not act in that way now, then it really makes no sense to say that I have such an intention now.

6.4 Degrees of Belief

One thing an account of belief must deal with is the question of degrees of confidence. An account that just allows for beliefs to be on or off is inadequate to the facts. This is a complex and interesting topic which I will do little justice to. As before, I just sketch how my account might be used to cope with the problems. Any gaps should just be attributed to the growing list of 'areas for future research'.

I propose the following analysis:

> *I believe with degree of confidence p that X if I am run by a partial teleological mechanism that gives results that would be rational were X to have probability p.*

I am simply assuming that there is a sensible notion of probability relative to the evidence. It seems clear that we do have ways of determining such probabilities, both theoretically and statistically. This is all that is required for me. Note that this notion of probability with respect to the evidence need not be absolute, so long as it is not subjective. There might be alternative ways of determining such probabilities. This would just mean that there were alternative teleological theoretical methods employing this notion.

Also, there is no reason for the determination of probability to have to produce exact figures. A degree of vagueness is positively to be recommended.

In support of these claims about probability, evidence and degrees of belief, consider an example. Suppose that I am standing at a T-junction and say that I am only 50% sure that our destination lies to the left. My companion tells me I am wrong to be only 50% sure. She is not making a judgement about how I am computing my internal evidence; how could she be? So what is she doing?

It seems very natural to think that she is claiming that my belief about the probability of the destination being to the left is false. She might be doing this on several different levels. She might mean that since the destination is in fact to the left I should be 100% confident of that. But this would fail to connect with my belief. My belief concerns the probability relative to the readily available evidence. That is why on this interpretation, her comment would be more of a joke than a claim that my belief was false. Suppose instead that she means that because if we look around us, we can see the familiar church spire of our destination to the left, the probability that our destination lies to the left relative to the available evidence is much higher than even. This might well satisfy me that I was wrong. But suppose instead that I say that I was working with a notion of probability relative to the evidence of my senses rather than the readily available evidence. This already sounds disingenuous. But then she persuades me that my eyes must have glanced over the church

spire, and that I failed to take in what it was that I was being presented with. If I still claim that my degree of confidence was quite correct relative to my internally available evidence, she will throw up her hands in despair. I am failing to communicate with her. Either I am deliberately winding her up, or I have a psychopathic inability to accept that I can make mistakes.

It is interesting to compare my account with Ramsey's in 'Truth and Probability' (Ramsey 1926). The similarity is that Ramsey regards degree of belief as being a function of how the agent is disposed to act. He tries to account for it in terms of objective tests of preference, rather than as something that can only be measured introspectively.

> *'Degree of belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it.'*
> (Ramsey 1926 p71)

There is an implication in the phrase 'degree of belief is a causal property *of it*' that I think is quite mistaken. This is that a belief is an entity independently identified, and only after it is identified as an entity is its degree determined in terms of its causal properties. The consequence of this would be that a proposition on which one had no views and so had a degree of belief of ½ is nevertheless believed. On my view, the degree of belief is not a property of a belief. Having a degree of belief is a property of the agent quite different from the property of having a belief.

But the major difference between Ramsey's account and my own is that Ramsey thinks he can do the job without reference to any notion of objective probability. He argues that the notion of degree of belief that he constructs is the only possible basis for probability theory. Now, indeed Ramsey makes no explicit reference to objective probabilities in his account. But it is my contenion that as a result his tests for degree of belief are entirely unmotivated.

One can use betting as a simplification of Ramsey's test. The idea is that if you are forced into a situation where you have only two possible courses of action, to bet on p or to bet on not-p, your degree of belief that p is a straightforward function of the odds at which you would be indifferent between the two bets. (Questions of risk aversion and diminishing utility of money complicate the issue, but make no difference to the argument.)

My complaint is not with the test as such, but that an important question remains unanswered. That is, why should the result of this test be regarded as constituting an agent's degree of belief? If it is simply a definition of degree of belief, then degree of belief is just a measure of betting disposition. But clearly, it is a measure of a much more general disposition. I suggest that the answer to the question is that the results of the test indicate a general disposition to act in a way which would be ~~if~~ rational were the *actual* probability of p just that laid down in the odds of the bet. If the actual probability was that laid down in the odds, then it would be rational to be indifferent

between the two alternatives. This is *why* the betting test works as a way of determining the degree of belief.

The sort of story I have told about degrees of confidence is generally regarded by decision theorists to be absurdly naive. (See for example papers in Gardenfors and Sahlin 1988.) Perhaps it involves some grotesque scope error. I am in fact identifying

I believe with degree of certainty p that X

with

I believe that X has p degree of certainty.

The degree of certainty operator has shifted its scope from the belief to the proposition. This is the sort of shift that usually spells disaster for a philosophical theory. But, I think that the right place for the operator is working on the proposition and that the misleading scope shift has happened in our language and not in my analysis. Our language leads us mistakenly to think of beliefs as mental entities that can have degrees of strength - shine stronger or weaker. It is probably quite important to diagnose this way of thinking, but I shall make do with registering my oppostion to it.

If my account is mistaken, then counterexamples should be available to show this. I shall consider some briefly. Suppose that I am full of self-doubt. I go for a job interview which goes well. I know that the chances of my getting the job are high. Yet I have absolutely no confidence that I will get the job. There are two kinds of case here. In one case I am really confident that I will get the job and act accordingly, but at the same time am assailed by feelings of doubt and

failure. This sort of case presents no direct threat to my account (although it does raise an issue which I will consider later that my account seems to have nothing to say about feelings). But consider another case. Suppose that my self-doubt has an effect on my way of behaving, so that I behave in a way that would be rational if the chances were small that I was going to get the job. For example, I am willing to throw away the option of accepting an offer on this job in exchange for a 10% chance of getting another equivalent job.

In this case I think we should say that I believe that my chances of getting the original job are slight. I may believe that according to the available evidence my chances *should* be thought of as high, but nevertheless I do not think of them as high. This is just a case of irrationality. There is no particular problem about degree of belief with this example. It is just the problem of believing something when one knows that all the evidence suggests that it is not true. This sort of case will be dealt with later, in the section on irrationality.

Consider another example which suggests that my account fails to make the distinction between uncertainty and real probability. Suppose that there are two dice in a box; throwing one gives a probability of 1/2 of getting a 6, while throwing the other gives a probability of 1/6 of getting a 6. One of the dice is taken out of the box, and is about to be thrown. I know all this, but I do not know which die has been taken. If I work it out, I should end up with a degree of confidence of 1/3 that a 6 will be thrown. However, I do not believe that the actual

probability of getting a 6 is 1/3. I believe that it is either 1/2 or 1/6, but I do not know which.

I suggest that this proposed counterexample simply trades on an ambiguity between different notions of probability, rather than opening up a gulf between belief that the probability is p and degree of confidence of p in the belief. Probability is a function of the evidence provided by a situation. If one includes in the evidence which die has been taken from the box, then the probability is either 1/2 or 1/6. If one does not, then the probability is 1/3.

There is no difficulty in my account capturing these different senses. One of my beliefs is that, relative to the facts that I have been able to acquire, the probability of a 6 is 1/3. Another of my beliefs is that relative to the facts that are theoretically acquirable, but not necessarily acquired, the probability of a 6 is either 1/6 or 1/2. The difference between the two beliefs can be manifested in behaviour mechanisms. Were it not rational or practicallly possible to acquire any more information, then, given that the probability of getting a 6 relative to the acquired facts is 1/3, it would be rational to bet on any better odds than these. Were it rational to gain more information about which die had been picked, then, given that the probability of a 6 relative to this further information is either 1/6 or 1/2, it is rational to plan my strategy accordingly, perhaps making conditional bets.

6.5 Behaviourism and the Holism of the Mental

In this section I discuss how well my account can be described as behaviourist. Since behaviourism is widely thought to be discredited, I need to show that the arguments which purport to do this discrediting do not discredit my account.

Braithwaite's theory of belief has a behaviourist element, but also includes a mentalistic element which strikes me as being entirely spurious.

> *'My thesis is that "I believe one of the propositions P", ... means the conjuction of the two propositions: (1) I entertain P ... , and (2) I have a disposition to act as if P were true.'*
> (Braithwaite 1932 p30)

Ryle was less compromising.

> *'To talk of a person's mind is ... to talk of the person's abilities, liabilities and inclinations to do and undergo certain sorts of things, and of the doing and undergoing of these things in the ordinary world.'*
> (Ryle 1949 p199)

There is a familiar problem with such dispositional accounts when it comes to spelling them out. That is, they cannot be spelled out. First of all, an indefinite number of caveats must be included in a specification of a behavioural disposition, to cover against things going wrong. For example, a specification of the disposition to take an umbrella if I go out in the rain needs to include in the conditional part that my brain is working, my eyes are open and working, etc.

Also, the specification will include in the conditional part conditions that can only themselves be described mentalistically: that I do not forget, that I do not change my mind, that I realize it is raining, etc. So the programme to analyse mentalistic terms in terms of non-mentalistic behavioural dispositions seems doomed.

This seems particularly clear because no mental state can ever be correctly described as the disposition to perform just one kind of act or even a finite number of kinds of acts. My believing that it is raining is not a disposition[5] to take an umbrella if I go out; for this behaviour only manifests the belief if I also want to stay dry, believe that taking the umbrella is the best way to stay dry, etc. If I have other desires and beliefs, other kinds of act will manifest my belief that it is raining. If a want a free shower I may run out into the garden naked.

I think that the first problem about the indefinite length of a specification of a disposition is dealt with by having a robust realism about dispositions. In my account, talk of a mechanism pulls this trick. To have a behavioural disposition is to have one's behaviour resulting from a mechanism which characteristically has such and such results. The operational conditions of the mechanism need not be included in this specification.

The second problem is that behavioural reductions are bound to be circular. One cannot provide a dispositional analysis of belief without referring to a desire, and vice-versa. This is described as the holism of the mental.

> *'Suppose someone has a particular belief at a particular time: then there will be repercussions on what that person will do in various circumstances, both at this and at other times. But these repercussiohns will be present only because the person has other beliefs and desires. Moreover, if a person possesses a particular belief, that seems to have no consequences for his actions that are consequences independently of his desires. The same holds for desires* vis-a-vis *belief. It follows that, in so far as beliefs and desires are ascribed on the basis of the actions they have as consequences, they cannot be ascribed singly but only in whole sets.'*
> (Peacocke 1979 p3)

Behaviourists try to incorporate this holism by analysing *sets* of beliefs and desires in terms of behavioural dispositions. But I think that this is unnecessary once the practical rationality that is constitutive of these dispositional states is seen to be objective, not subjective.

What beliefs or intentions a person should be attributed with depends on what would make sense of that person's way of behaving. So a consideration of rationality partly determines what mental states we should attribute to them. On a subjective (or Humean) conception of rationality in which the rationality of beliefs and intentions depends on what other beliefs and intentions the person has, this means that the attribution of any beliefs or intentions depends on what other beliefs and intentions are attributed. Hence holism of the mental follows from thinking of practical rationality as essentially subjective.

One target of this argument is the conception of mental states as independently identifiable entities. I have no quarrel with this aspect

of the argument, although my view of what they are not independent of is different. The constitutive role of rationality is very much my starting point also. The only quarrel I have is with the assumption that it is subjective rationality that should determine attributions of mental states. Mental states are supposedly attributed to make sense of a way of behaving. But if the very rationality underlying this is itself subjective, no explanatory work is being done by the attribution. Why should the attribution of a belief that ϕ is the means to achieving ψ and a desire for ψ make sense of behaviour leading to ϕ. If the belief and desire are only attributed in virtue of making sense of this sort of behaviour, then there is no independent conception of what it is about beliefs and desires that does make sense of the behaviour.

Given this, we can see that functionalists are really cheating badly. For, they rely on the fact that we have a conception of beliefs and desires other than 'states that play certain functional roles in the explanation of behaviour' so that we are persuaded that beliefs and desires do have roles in the explanation of behaviour. If our only conception of what beliefs and desires were was the functionalist conception, then the attribution of beliefs and desires would entirely fail to make sense of behaviour, and the whole functionalist programme would collapse.

To make this clearer, let us change the words. If the functionalists are not making any use of a non-functionalist conception of beliefs and desires, let us instead talk of A-states and B-states whose meanings are entirely determined by their functional roles.

Suppose I wish to make sense of a person's behaviour by attributing states which I call 'A-states' and 'B-states'. The method is to work out a set of descriptions of the form, 'A-state that ϕ *is a means to* ψ is present', and, 'B-state that ψ is present'. The set should maximally satisfy the condition that pairs of descriptions, 'A-state that ϕ *is a means to* ψ is present', and, 'B-state that ψ is present', are associated with behavioural tendencies to achieve ϕ. (Other conditions can be included as well, but will not materially affect the argument.)

So I have attributed A-states and B-states according to a principle given the person's way of behaving. In what way does making this attribution *make sense* of the person's way of behaving? How can we suddenly say that the behaviour is *right* just because we have been able to apply the principle and yield the attribution? All we can say is that the behaviour is such as it is possible to apply the principle and yield the attribution. But this is not enough to make sense of the behaviour. We also need some reason to suppose that the application of the principle means that the behaviour makes sense. But for this we need some notion of rationality that gets us out of this subjective circle.

On my account, beliefs and intentions are not attributed in a way that fits in with other beliefs and intentions (that are being attributed in the same way ...). But they are attributed in a way that fits in with some way of objectively (i.e. really) rationalizing the behaviour. This way is the theoretical method representing the agent's complete teleological mechanism. Once this way is determined, then

beliefs and intentions can be attributed singly or in small groups, but not holistically. Of course, something very like holism is operating here. Beliefs and intentions are not attributed independently of some overall explanatory structure (the complete teleological theoretical method). But what is crucially unholistic is that this structure can be attributed independently of any beliefs or intentions the agent may have.

Now, of course, a single piece of behaviour is not enough to justify an attribution of a single belief or desire. By itself it can only justify accepting a disjunction of alternative sets of beliefs and desires. If one considers that the correct attribution of beliefs and desires is determined by a process of working through pieces of behaviour one at a time until a unique set is found, then the holistic conclusion is inevitable. Indeed in practice this *is* roughly what we do. My objection is to the claim that this process *constitutes* the correct attribution. For, on my account, the sort of rationality that the correct attribution depends on constitutively does not consist in the interaction of beliefs and desires.

Accepting that we might have to know about what behaviour would ensue in indefinitely many possible situations in order to attribute the correct beliefs and desires to an individual, the pressure towards constitutive holism recedes. What the underlying complete teleological theoretical method is is not dependent on any attribution of beliefs or intentions. It is determined by considering the behaviour in all situations where the relevant information is apparent. So to determine

the structure of goals according to which the complete teleological mechanism operates, no consideration of beliefs is necessary. One need only consider situations where the truth is so apparent that if the individual is rational at all they will act on it. Then having derived this system of goals, one can consider situations where the agent must act without getting any further information concerning the truth or rationality of certain propositions or courses of action. Thus beliefs and intentions are determined.

One further argument against a Ramsey-style behaviourism is Davidson's. (See Davidson 1974 pp 141-154 and 1990.) His argument is that one cannot attribute beliefs and intentions to an agent on the basis on non-linguistic behaviour only. It is not reasonable to attribute to me the belief that a squirrel ran up the tree rather than the belief that some furry animal ran up the tree unless I am disposed to *describe* the event as that of a squirrel running up the tree. But linguistic behaviour itself needs to be interpreted in terms of the agent's beliefs and intentions. So, according to Davidson, there is another dimension to the holism - the dimension of the agent's meanings. What grounds this holistic structure is the interpreter themself through the principle of charity.

There is a fairly trivial sense in which it is true that the interpretation is up to the interpreter. This is not a threat to my account. It is not that the way that a mechanism works is up to the interpreter. The way that a mechanism works is fixed to that mechanism. But what mechanism is picked out from a choice of

metaphysically dependent mechanisms is up to the interpreter. There need be no unique ultimately objective way of describing things.

But Davidson is making a further point; that there is a special problem with the interpretation of intentional behaviour. His thought is that there is something especially indeterminate about the interpretation of an agent's meanings; this interpretation relies on being able to interpret the agent's linguistic behaviour; and this itself relies on a very peculiar assumption being made by the interpreter. That assumption is the principle of charity, that most of the agent's beliefs and values coincide with the interpreter's own.

I am not convinced that we need linguistic behaviour in order to attribute beliefs and intentions. But this is not the crux as I see it. The point about the principle of charity would apply even if the agent did not use language. But my account incorporates the intuition behind this principle. For, it is an essential aspect of a teleological mechanism that when it works its results are objectively correct. When a fact is absolutely apparent to a teleological mechanism whose operational conditions are generally satisfied, we can assume that in respect of that fact its operational conditions are satisfied also. So we can assume that when a fact is obvious the teleological mechanism will get that fact right. This assumption is not a peculiar interpretational principle. It is part of our conception of what a teleological mechanism is. So the principle of charity is an aspect of the teleological mechanism; it is not an extra interpretational condition required to cope with irreducible holism about meanings.

The question of how my account intersects with questions of content and the philosophy of language is clearly very important, but will not be dealt with in this thesis. The main thing to realize is that my account explains how we attribute one content rather than another to an individual. It is on the basis of whether that individual can discriminate in their way of behaving at the level of one content or at the level of the other. If some kinds of furry animals are edible and others are not (and this fact is apparent to me), then whether I just believe that a furry animal has gone up the tree or whether I also believe that an edible furry animal has gone up the tree can be established by considering whether I try and catch it in a situation where this is possible and I am hungry.

Whether or not language is essential for discriminating any contents, it is certainly one way it is done. If I am a language user, then I may believe that a squirrel has gone up the tree even though I cannot distinguish a squirrel from a chipmunk. For my way of behaving with respect to the language community (i.e. acting as if the thing these people call a squirrel has gone up the tree) does in fact discriminate between squirrels and chipmunks. This captures Putnam's notion of the linguistic division of labour. So it looks as if this account should be able to incorporate thoughts about broad content. (c.f. Pettit and McDowell 1986.)

Because it is not explicitly holistic, my account involves a sort of reductionism. This needs to be clarified. My theory suggests that talk of an agent's mental states can be reduced to talk of their

partial teleological mechanisms. This is a sort of behaviourist reduction, but it is very far from reducing mental language to purely 'physical' or non-intentional terms. There are two reasons for this. One is that the way the mechanism works is specified in intentional terms. The other is that rationality is an essential part of how the mechanism works. One cannot reduce away reference to rationality or to a theoretical method embodying rationality. (This is Davidson's point about the anomalousness of the mental deriving from the 'constitutive ideal of rationality' in Davidson 1970.)

I am not even committed to an account identifying mental states with physical states. A large degree of ontological liberality was an essential part of my treatment of mechanisms. I do not have to be committed to any form of dualism or pluralism about *substances* to deny that an agent is a body or mental states are bodily states.

What emerges from my account is a kind of hierarchy of mental concepts. First of all there are agency and purpose. These are to be understood in terms of the working of an objective teleological mechanism. Whenever a process can be teleologically explained in terms of that process being a means to an end, then the system whose working results in the process is an agent and the end of that process is the purpose of the process.

Establishing the facts can be identified as part of that process. Then perception, reasoning and action can be identified in terms of that. Then there are beliefs and intentions. By no means can every aspect of the behaviour of an agent be explained just in terms of the

complete teleological mechanism. So we identify partial teleological mechanisms essentially operating on sub-optimal amounts of information. Using these we can explain behaviour that does not conform to the result of the complete teleological mechanism. What beliefs and intentions an agent has are determined by what partial teleological mechanisms are operating.

Can this strategy be extended? Are there aspects of an agent's behaviour that need further types of adaptation of the basic teleological mechanism to explain? And can further mental states be understood in terms of the presence of such adapted teleological mechanisms? Are all mental states to be so understood? My view is that the answer to each one of these quesions is yes. For example, it may be possible to distinguish between long- and short-term teleological mechanisms. Where a long-term theoretical method is indeterminate, short-term teleological mechanisms may encapsulate things like idle curiosity, just feeling like doing something, etc. But these considerations really will have to be relegated to the category of 'future research'.

6.6 Rational and Causal Roles of Beliefs and Intentions

It has been no part of my aim to deny that beliefs and intentions have rational and causal roles. My aim has been to show that these roles are not analytically fundamental. It might be thought that I have nevertheless left myself no room for the notion of subjective rationality, whether theoretical or practical. So I shall briefly indicate how to construct the idea of subjective rationality from my account. In doing this I will show that the only reason subjective rationality works at all is that objective rationality underlies it.

What kind of reasons are subjective reasons? They are not purely explanatory; so are they justificatory as well? It would seem so at first. But if they are justificatory, I do not think that it is in the sense that I analysed earlier; that is having an essential role in a system of practical rationality embodying means and ends.

Consider how subjective reasons might play such roles. Either the ends are external or partially external; for example, to act well or to have a high proportion of true beliefs. In such cases it makes no sense to think of the reasons as being completely internal; external factors will be part of the justificatory story. Or, the ends are internal; for example, to have internally founded beliefs and intentions or to have internally consistent beliefs and intentions. This must be quite close to what a believer in the importance of subjective rationality has in mind.

The problen is that such ends hardly make much sense as ends in themselves. Surely there must be some further reason why I would want to achieve such ends. Indeed, in practice, nobody has either of these aims as top-level goals. There will always be situations where it makes good sense to keep an inconsistent set of beliefs. And even the very possiblity of having internally well founded beliefs and intentions is open to serious doubt.

So it seems that subjective rationality is not really justificatory. I suggest that subjective reasons provide *excuses* rather than real justifications. They only justify in the sense of 'mitigate'. The basis for this suggestion is in the distinction between a reason *meaning* it is rational to ϕ and a reason *making* it rational to ϕ. Truly justificatory reasons make what they justify rational. Mitigatory reasons only show what they justify to be rational.

Intending to eat means that it is rational to eat, but it does not ***make*** it rational. Believing that the food is poisonous means that it is rational not to eat, but it does not ***make*** it rational. What makes it rational is the *fact* that the food is poisonous or the fact that the evidence suggests that the food is poisonous. Similarly, believing the food to be poisonous means that it is rational to believe that I will be unwell if I eat it. But again, it does not ***make*** this belief rational. What makes the belief rational are the *facts* on which the belief is based.

So my analysis of a mitigatory reason is the following:

X is a mitigatory reason for Y if and only if X is rational given Y.

If a belief that p is a mitigatory reason for the belief that q, this means that it is rational to believe that q given the belief that p. A rational agent who already believed that p would thereby believe that q.

Suppose that I believe that ~~q~~ p and I believe that if-p-then-q, and suppose that my teleological mechanism is working fine, at least after the acquisition of these beliefs. It follows that I am operating with a mechanism which results in what would be rational were p and if-p-then-q both true in situations where no more establishing of the facts concerning the truth of p or of if-p-then-q occurs. If I already believe that not-q, then it will not necessarily be rational to believe that q; it may be rational to believe that not-p instead. But suppose that I do not already believe that not-q. Then in situations where there is no more establishing of the facts concerning p or if-p-then-q, and my teleological mechanism works, then I will do what would be rational were q true. Does it follow that if there is no more establishing of the facts concerning the truth of q, I will do what would be rational if q were true? Yes. Any establishing of whether p or if-p-then-q are true is establishing of whether q is true, and vice-versa. So, if my teleological mechanism is working properly after the acquisition of the beliefs of p and of if-p-then-q, and if I do not believe that not-q, I will thereby believe that q.

Exactly the same argument shows that beliefs and intentions may be mitigatory reasons for new intentions, and new intentions may be mitigatory reasons for actions.

If this is right, then Humean subjective rationality has very little philosophical importance. It is entirely derivative on objective practical rationality. In terms of its role in causal explanation it has very little usefulness either. By this I do not mean that thoughts and feelings do not have causal roles. What I reject is the idea that there is a mechanism in a person which takes beliefs and intentions as inputs and results in more beliefs and intentions as well as behaviour in a way that corresponds with subjective rationality. Having a belief or an intention is having a certain sort of partial teleological mechanism; it is not an input to such a mechanism. So when beliefs and intentions are referred to in the explanation of some piece of behaviour, what is being specified is the mechanism resulting in the behaviour, not inputs to that mechanism. The background to the causal explanation is being described, rather than the causes themselves.

An argument in support of this is that subjective reasons, though sometimes leading causally to the things they mitigate, very often do not. For example, forming the belief that I had received no mail this morning, when I went and checked, did not cause me to have the belief that the letter from Smith had not come yet. The two beliefs arose simultaneoulsy, even though one may be said to have rationalized the other.

On my account, having a belief is the same as being run by a partial teleological mechanism. So having a belief should result in various things through the process of this mechanism working. This is just what happens. My teleological mechanism embodying the belief that I have not received Smith's letter may result, through the process of perception, in the formation of other beliefs; e.g. I ring up Smith and discover that she has gone on holiday. Or the process of reasoning might result in my coming to believe that Smith does not care as much about me as she claims, since she cannot get the letter to me in time. Or the process of action might result in me writing Smith another letter in case the first one had not got through. Each of these processes is part of the larger process of my partial teleological mechanism, which embodies my belief about Smith, working.

6.7 Irrationality and Acrasia

Any theory of intentional mental states that gives rationality a constitutive role has to explain how irrationality of beliefs and intentions is even logically possible. The first sort of irrationality to consider is that of believing logically impossible propositions. Suppose, for example, that I believe that $8^2=36$. There is an apparent problem with this on my account. For this should mean that I do what would be rational were $8^2=36$. But from the proposition $8^2=36$ every other proposition, true or false, follows. (This claim might depend on

one's philosophy of mathematics, but in that case an example of a purely logical contradiction will do just as well.) This means that the theoretical method breaks down and any behaviour is rational according to it.

One might try to retreat from the brink of this catastrophic result by claiming that the belief is really about names not numbers. But I think that a more promising strategy is to deny the account of counterfactuals underlying the argument. Clearly, every proposition follows logically from a contradiction. But this does not mean that were the contradiction true, every other proposition would be true also. It would mean this or something like it if the 'possible worlds' analysis of counterfactual conditionals were correct. Possible worlds are generally thought to incorporate the rules of logic. So, there is no possible world in which $8^2=36$. This means that on a possible worlds account it would be impossible to believe that $8^2=36$. We could not make sense of a theoretical method working as if $8^2=36$, since we cannot make sense of the possiblity that $8^2=36$.

However there are other ways of understanding counterfactuals. For example, consider the 'make-believe' account. (See Evans 1982 pp353-363 who takes the idea from Walton.) The truth of a counterfactual claim may be considered to be relative to the rules of the make-believe game that is being used. Suppose (i.e. pretend) that $8^2=36$ were true, then it would be rational to say so to the maths teacher (in certain circumstances). It would also be rational to substitute 8^2 for 36 in equations. If we can make sense of this supposition at all, then we

are not using a possible worlds account where the laws of logic are incorporated in every possible world.

The difficulty may be thought to recur however. For, if one were faced with the equation $6^2=36$, it would be rational to make the substitution of 8^2 for 36 and assert that $6^2=8^2$, and then it would be rational to assert that 6=8. The problem is that it does not follow from the fact that I believe that $8^2=36$ that I believe that 6=8.

However my account copes with this perfectly. Believing that $8^2=36$, I am not supposed to do what would be rational in *every* situation where $8^2=36$. I am only supposed to do what would be rational in every situation where no more perception or reasoning (establishing of the facts) concerning that proposition occur. The situation just described involves reasoning leading to the conclusion that 8^2 is not equal to 36. So it is ruled out of consideration as a possible input situation for the partial teleological mechanism embodying the belief that $8^2=36$. Note, that it is *not* ruled out of consideration as a possible input situation for the complete teleological mechanism. So, as long as there is nothing wrong with the way that is working, that piece of reasoning will occur, and I will no longer act as if $8^2=36$. So the belief is basically unstable, although it is not impossible.

This is why it is possible to hold an inconsistent triad of beliefs. For example, suppose I believe that I will be abroad for the second half of August on holiday. Suppose also that I absent-mindedly arrange to meet someone at home on August 20th, and I believe that I

will keep the appointment. And suppose finally that I know that these two things are incompatible. On my account, the fact that I believe these three propositions just is the fact that I am run by a teleological mechanism, that in situations where no establishing of contrary facts occurs, results in behaviour that would be rational were it the case that I will be abroad in the second half of August, at home on August 20th and these two things mutually incompatible.

No problems emerge where the rational course of action depends on just one or two of these inconsistent propositions. But what about situations where the inconsistency bites. For example, suppose that I need to tell someone where I will be on August 20th. I cannot do what would be rational were all three propositions true. But as before, this is not a problem, since this situation is not one that the partial mechanism embodying these three beliefs needs to deal with. For the situation essentially involves reasoning which establishes the falsity of one of the propostitions.

This solution to the apparent problems of this kind of irrationality does not work against the apparent problem raised by the possibility of believing a pair of directly contradictory propositions p and not-p. For there is no situation where a partial teleological mechanism could have results that were rational were p and not-p both true. It is an entirely redundant specification of the partial teleological mechanism to incorporate the beliefs that p and not-p, because there is no situation where this aspect of the specification makes any difference to what would result if the mechanism worked.

This is desirable in a way, since it does seem to be absurd to attribute anyone with beliefs in both a proposition and its contrary. But there are situations where we feel it might be appropriate to make such attribution. So let us consider them.

First of all, I might have wavering beliefs. I might believe p sometimes and believe not-p at other times. This raises no particularly difficult problems. Similarly, if my beliefs that p and that not-p are only partial (i.e. their degrees of belief are less than one), there is no problem about their joint attribution.

Problems for my account arise if I may be in a state in which in certain situations I do what would be rational were p true, and in other situations I do what would be rational were not-p true. According to my account, we should say that I do not believe either p or not-p. However, it may be more natural to claim that in some sense I believe them both. Perhaps, deep down I believe that p, but in terms of my more superficial behaviour patterns I act as if not-p were true.

One way that this might work is if I really believe that p, but also believe that I should believe that not-p. The belief that I should believe that not-p manifests itself in very similar ways to the ways that the simple belief that not-p manifests itself. For example, in situations involving communication, it may be rational to assert not-p rather than p. My purpose in communicating may be to impress or to conform rather than to tell the truth.

If my belief that I should believe not-p is the belief that, according to the generally accepted theoretical method, it is rational

to believe that not-p, then there is no actual irrationality here. A more interesting case is where I believe that, according to *my own theoretical method*, I should believe that not-p, nevertheless I believe that p. One way that this might happen is if my belief about the workings of my own theoretical method is false. For example, I may believe that I share the beliefs of my twin, and for this reason come to believe that according to my theoretical method I should believe that not-p, nevertheless I believe that p. This case is equivalent to having a false belief about my beliefs. It is an unstable position to be in, but it is a situation quite consistent with my account.

But consider a more difficult case. Suppose that I am presented with the information that I have an Oedipal complex, but I do not believe it, because then I would have to accept that I am not the sort of person I wanted to be. This reveals a kind of irrationality that looks like a real problem for my account. Assume the operational conditions of my complete teleological mechanism are satisfied, and that I am clearly presented with the information that I do have an Oedipal complex. Then, behaviour whose rationality depends on this fact must result given that I am run by a teleological mechanism. Yet, if I have a mechanism that results in behaviour that would be rational if I had an Oedipal complex, I believe I have an Oedipal complex. Contradiction.

This sort of case (following Pears 1984) can be described as motivated irrationality. I think that the right way to interpret it is to deny the step between the following positions:

A. If the truth of p makes some piece of behaviour rational then that behaviour will occur.

B. A mechanism is present which results in behaviour which would be rational were p true.

In the case of motivated irrationality the agent makes sure that in no situations does the rationality of their behaviour depend on the truth of p. In the example under consideration, I am not in fact motivated by the goal of speaking the truth about myself; so it is quite rational for me to deny that I have an Oedipal complex. It remains true that if the rationality of some behaviour depended on the fact that I have an Oedipal complex, that behaviour would have to result if any teleological mechanism was operating. But, I make sure that my structure of goals is such that no such behaviour could be required. So it is not the case that I am run by a mechanism that results in what would be rational if I had an Oedipal complex. There is no situation where that aspect of the supposed mechanism could manifest itself. So I do not believe I have an Oedipal complex.

The restructuring of goals necessary to avoid having this belief may well leave me with inconsistencies in my beliefs. But they will be of the unproblematic kind rather than a straight face off between p and not-p. It may also make it difficult for me to build up structures of goals. It is a commonly observed psychological fact that people with motivated irrationality may often fail to make very much of themselves. In the extreme case I will be rendered incapable of doing almost anything at all, because avoiding having the belief about myself will

be impossible unless I do nothing. (Note that good psychological therapy is more concerned with undermining the motivation for not believing that p than with battering away at the client with the fact that p.)

Now consider acrasia and irrationality with intentions. There are two kinds to be considered. In the first kind the agent intends to ϕ but fails to ϕ. In the second the agent believes that they should not ϕ, but nevertheless they intentionally ϕ.

The first kind of acrasia is not problematic. If something goes wrong with the teleological mechanism after the intention has formed or the mechanism simply changes before the final behaviour is due, then a formed intention might fail to lead to the appropriate behaviour. What would be a problem would be the possibility of doing ϕ at the same time as intending not to. But this is not a possibility.

The second sort of case is more interesting. Suppose that I believe that I should not ϕ, yet I intend to ϕ. Intending to ϕ involves being run by a partial teleological mechanism that results in what would be rational if I should not ϕ. So we have exactly the same problems we had with accounting for apparently contradictory beliefs. And the solutions are the same as well.

One interpretation is that the 'should' in my belief that I should not ϕ is relative to some theoretical method not my own, for example society's or the theoretical method that I should have. Another interpretation is that the 'should' is relative to my own theoretical method, but my belief about my theoretical method is false. This may

be because of some natural failure of perception or reasoning. Or it may be bound up with motivated irrationality. The same analysis works here as was applied to motivatedly irrational beliefs. I do not really believe that I should not ϕ, and if I have to undermine my whole structure of goals to maintain this belief, I will.

References

Achinstein, P. (1977). "Function Statements", *Philosophy of Science*, 44.

Achinstein, P. (1983). *The Nature of Explanation*. Oxford: OUP.

Anscombe, G.E.M. (1971). *Causality and Determination*. Cambridge: CUP.

Annas, J. (1982). "Aristotle's Inefficient Causes", *Phil. Q. 1982*.

Bennett, J. (1976). *Linguistic Behaviour*. Cambridge: CUP.

Blackburn, S. (1973). *Reason and Prediction*. Cambridge: CUP.

Blackburn, S. (1984). *Spreading the Word*. Oxford: Clarendon Press.

Boorse, C. (1976). "Wright on Functions", *Phil. Rev. 85*.

Braithwaite, R.B. (1932). "The Nature of Believing", *PAS 33*, reprinted in Phillips Griffiths (1967).

Braithwaite, R.B. (1953). *Scientific Explanation*. London: CUP.

Bratman, M. (1987). *Intentions, Plans and Practical Reasoning*. Cambridge: Harvard UP.

Bunge, M. (1959). *Causality*. Cambridge: Harvard UP.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, N. (1989). *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Charles, D. (1984). *Aristotle's Philosophy of Action*. London: Duckworth.

Charles, D. (1989). "Intention", in Heil (ed), *Essays in Honour of C.B. Martin*.

Charles, D. (forthcoming). "Teleological Causation in the Physics", in Judson (ed).

Chisholm, R.M. (1966). *Freedom and Action* , in Lehrer 1966.

Clark, A. (1989). *Microcognition*. Cambridge: MIT Press.

Cooper, J. (1987). "Hypothetical Necessity and Natural Teleology", in Gotthelf and Lennox (1987).

Cummins, R. (1985). *The Nature of Scientific Explanation*. Cambridge: MIT Press.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, D. (1990). "The Structure and Content of Truth", *J.Phil* 87: 6.

Davies, M. (1985). "Function in Perception", *Australasian Journal of Philosophy* vol 61 no 4.

Dawkins, R. (1976). *The Selfish Gene*. Oxford: OUP.

Dennett, D. (1984). *Elbow Room*. Oxford: Clarendon Press.

Dennett, D. (1987). *The Intentional Stance*. Cambridge: MIT Press.

Dretske, F. (1988). *Explaining Behaviour*. Cambridge: MIT Press.

Dummett, M. (1978). *Truth and Other Enigmas*. London: Duckworth.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Clarendon Press.

Evans, G. (1985). *Collected Papers*. Oxford: Clarendon Press.

Fodor, J. (1981). *Representations*. Cambridge: MIT Press.

Frankfurt, H. (1971). "Freedom of the Will and the Concept of a Person," *J.Phil* Jan 1971. (Reprinted in Watson 1982.)

Gardenfors, P. and Sahlin, N-E. eds (1988). *Decision, Probability and Utility*. Cambridge: CUP.

Gibson, J.J. (1966). *The Senses considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gotthelf, A. (1987). "Aristotle's Conception of Final Causation", in Gotthelf and Lennox (1987).

Gotthelf, A. and Lennox, G. eds (1987). *Philosophical Issues in Artistotle's Biology*. Cambridge: CUP.

Hinton, J.M. (1973). *Experiences*. Oxford: Clarendon Press.

Hornsby, J. (1980). *Actions*. London: Routledge, Kegan & Paul.

Hurley, S. (1989). *Natural Reasons, Personality and Polity*. Oxford: OUP.

Kenny, A. (1975). *Will, Freedom and Action*. Oxford: Blackwell.

Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.

Lepore, E. & McLaughlin, B., eds (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Clarendon Press.

Lewontin,R. (1983). "Darwin's Revolution", *New York Rev. of Books*, 30: 21-27.

Lloyd, D. (1989). *Simple Minds*. Cambridge: MIT Press.

Loar, B. (1981). *Mind and Meaning*. Cambridge: CUP.

McDowell, J. (1978). "Are Moral Requirements Hypothetical Inperatives?" *Principles of the Aristotelian Society, Supplementary Vol* 1978.

McDowell, J. (1979). "Virtue and Reason," *Monist* vol 62 no 3.

McDowell, J. (1982). "Criteria, Defeasibility and Knowledge", *Proc B.A.* OUP.

McDowell, J. (1985). "Functionalism and Anomolous Monism", in Lepore and McLaughlin (1985).

McDowell, J. (1986). "Singular Thought and the Extent of Inner Space", in Pettit and McDowell eds (1986).

Melden, A.I. (1961). *Free Action*. London: Routledge, Kegan & Paul.

Millikan, R. (1985). *Language, Thought and Other Biological Categories*. Cambridge: MIT Press.

Nagel, E. (1961). *The Structure of Science*. Indianapolis: Hackett.

Nagel, T. (1970). *The Possibility of Altruism*. Oxford: Clarendon Press.

O'Shaugnessy, B, (1980). *The Will*. Cambridge: CUP.

Peacocke, C. (1979). *Holistic Explanation*. Oxford: Clarendon Press.

Peacocke, C. (1985). *Sense and Content*. Oxford: Clarendon Press.

Pears, D. (1975). "The Appropriate Causation of Basic Intentional Actions," *Critica* VII no 20.

Pears, D. (1984). *Motivated Irrationality*. Oxford: Clarendon Press.

Pettit, P. & McDowell, J. eds (1986). *Subject, Thought and Context*. Oxford: Clarendon Press.

Phillips Griffiths, A. ed (1967). *Knowledge and Belief*. Oxford: OUP.

Pitcher (1971). *A Theory of Perception*. Princeton: UP.

Putnam, H. (1988). *Representation and Reality*. Cambridge: MIT Press.

Ramsey, F. (1926). "Truth and Probability", in *Foundations* (1978), London: Routledge, Kegan & Paul.

Rescher, N. (1970). *Scientific Explanation*. New York: Free Press.

Rumelhart, D. and McLelland, J. eds (1986). *Parallel Distributed Processing*. Cambridge: MIT Press.

Ryle, G. (1949). *The Concept of a Person*. Hutchinson.

Searle, J. (1985). *Intentionality*. Cambridge: CUP.

Sober, E. (1984). *The Nature of Selection*. Cambridge: MIT Press.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*.

Stoutland, F. (1976). "Causation and Behaviour", in *Acta Philosophica Fennica 28: Essays in Honour of G.H. Von Wright*.

Strawson, P. (1959). *Individuals*. London: Methuen.

Taylor, C. (1964). *The Explanation of Behaviour*. London: Routledge, Kegan & Paul.

Taylor, C. (1985). *Human Agency and Language - Philosophical Papers 1*. Cambridge University Press.

Taylor, R. (1966). *Action and Purpose*. New Jersey: Prentice Hall

Watson, G. ed (1982). *Free Will*. Oxford University Press.

Wiggins, D. (1980). *Sameness and Substance*. Oxford: Blackwell.

Williams, B. (1980). "Internal and External Reasons", in Williams (1981), *Moral Luck*. Cambridge: CUP.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford: Blackwell.

Woodfield, A. (1976). *Teleology*. Cambridge: CUP.

Wright, G.H. von (1971). *Explanation and Understanding*. Lolndon: Routledge and Kegan Paul.

Wright, L. (1973). "Functions", *Phil Rev*. 82.

Wright, L. (1976). *Teleological Explanation*. Berkeley: UP.

Wynne-Edwards, V.C. (1962). *Animal Dispersion in Relation to Social Behaviour*. Edinburgh: Oliver and Boyd.