Dennett's Compatibilism Considered

THESIS

Submitted in partial fulfilment
of the requirements for the Degree of
Master of Arts
of Rhodes University

BY

Julian Gatenby Puttergill

January 1997

Acknowledgements:

Many people made my writing of this thesis possible and I would like to take this opportunity to express my gratitude to them. So, thanks to Francis Williamson who helped me to understand my thoughts and got me to believe in something, to the Rhodes University Philosophy Department which invested so much time and patience in my training, and finally to Marsh for digging my heels out every time I stuck them in!

The financial assistance of the Centre for Science Development (HSRC, South Africa) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the Centre for Science Development.

Abstract - Dennett's Compatibilism Considered

My basic concern in this thesis is to examine the details behind Dennett's attempt to reconcile the notions of mechanism and responsibility. In the main this involves an examination of how he tries to secure a compatibilism between mechanistic and intentional explanations by developing a systematised conception of intentional explanation.

I begin by briefly discussing the various notions needed for understanding what is at stake in the area and where the orthodoxy on the matter lies. As such the first three sections of the work are not focussed on Dennett's work itself and play a stage-setting role for the deeper work to follow. These notions include the likes of the rationale behind attributing moral responsibility, agency and action, mechanism and mechanistic explanation, and intentional explanation. I suggest that the basic intuition regarding mechanism and responsibility is such that the two are seen to be incompatible with each other. The main reason for this lies in an intuition that mechanism undermines intentional explanation and so renders the notion of action largely empty. Action, I show, is at the heart of our attribution of responsibility and is dependent on intentional explanation. Having presented these issues, I turn to the details of Dennett's 'intentional systems theory'.

I argue that Dennett attempts to avoid the intuition that mechanism is incompatible with responsibility by developing a specialised account of intentional explanation. Dennett calls it the *intentional stance*. I highlight the two important features of this *intentional stance*, namely rationality and intentionality. I show that Dennett's position on rationality and intentionality is such that it *does* allow him to secure an explanatory compatibilism between mechanism and his sort of intentional explanation. I argue, however, that his sort of intentional explanation *does not* fulfil our requirements for ascribing agency or moral responsibility. This is accomplished in part by developing alternative conceptions of the two notions. Out of this I develop a different sort of *intentional stance*, which I call the *folk stance*. I show finally that Dennett's compatibilist move is incapable of being applied to the *folk stance* from which we do *in fact* make attributions of responsibility, and so conclude that Dennett fails to make the case for reconciling mechanism and responsibility.

Table of Contents

Illustrations

1.0 Introduction

In Dennett's paper, *Mechanism and Responsibility*, he attempts to show that intentional explanations are compatible with mechanistic explanations and thus that mechanism and moral responsibility are also compatible. The basic idea involved is that moral responsibility is an ascription or a judgement that pertains strictly to the domain of action. Actions, of course, are instances of behaviour that are intentionally explained. So, if intentional explanation can be reconciled with mechanistic explanation then as long as we can provide an intentional explanation of a mechanistic system it can be said to act and thus be held morally responsible. Dennett accepts that we do not *actually* hold all such mechanistic systems responsible and even that we do not *always* consider them as agents of the moral community. When we do hold such systems responsible we have, according to Dennett, *decided* to consider the systems as being persons. The fact is, however, that as far as Dennett is concerned, many if not all systems that we consider to be persons and hold morally accountable may in fact be mechanistic in nature and he seeks to show that if this is the case the mechanistic nature of the systems does not impact on the issue of their being responsible. This position is typically called compatibilism. It is contrasted with what can be called incompatibilism or libertarianism.

Libertarianism is basically the idea that people can only be held morally responsible for those things for which they are also causally responsible. In other words, we can only be held responsible for the things we actually *do* and not the things that *happen* to us or for which we are merely the instruments. Now obviously when we are faced with a mechanistic system such as a thermostat, for example, we cannot, on this account, say that it is causally responsible for what it does. Certain antecedent environmental events are causally responsible for its behaviour and so it could not be a candidate for moral responsibility. What this position suggests is that there is a deep incompatibility between the notions of moral responsibility and causal necessity. This incompatibility is sometimes called *modal incompatibility*[1] because it involves the idea that the logical modality of necessity is just completely incompatible with the freedom required for responsibility. To say the same thing in simpler terms, there is no point at which the two types of discourse can meet. Trying to make them do so would be the

---

[1]The term is coined, I suspect, by Gary Watson. See his book *Free Will*, p. 12 for a discussion.

equivalent of committing a category error. I believe that this libertarian position is basically the intuitive position with regard to responsibility. When we say that someone or something is morally responsible we *want* to be saying that ultimately the buck must stop with that system. There is small satisfaction in the idea that we blame a system for what it does but do not really have to because there is a story we can tell which cites antecedent events beyond the control of the system as the causal features of the behaviour. In essence then, the intuitive perspective is one where we feel that we can only legitimately assign moral responsibility in cases where we have no choice but to explain the system's behaviour in terms of its being causally responsible for what it does. Explanations which render the system itself causally responsible are, obviously, intentional explanations.

It is clear that to respond to the intuition that there is a modal incompatibility between the notions of mechanism and responsibility one has to begin by showing that as explanations, mechanistic and intentional accounts are not mutually exclusive. What has to be shown is that the two types of explanation are not competitors as such and can both be given equal credence even if applied to the same instance of behaviour. This is what Dennett seeks to do, and can be called an attempt to establish *explanatory compatibilism*. What is clear is that his position is vastly different from the intuitive position and thus deserves attention. He is advocating a shift in our normal approach and thinking on this matter and so deserves due consideration and scrutiny. I believe that Dennett is, in the final analysis, unsuccessful in achieving his aim. What I propose to do is to show why I believe this. Basically I investigate the details of both our intuitions and Dennett's account and show that, valuable as his account is, he fails to persuade us to abandon our intuitive position.

Structurally the investigation takes the following form. I begin with a section which seeks to establish with clarity the notion that the ascription of moral responsibility requires at least being able to provide a purposive or intentional explanation of behaviour. I rely here on an intuitive or common sense understanding of both intentional explanation and the ascription of moral responsibility. In the next section I suggest, from an intuitive perspective again, that there is a strong sense in which intentional explanation might be thought to be undermined by the truth of mechanism. At this stage I present a comprehensive explanation of the notion of mechanism and mechanistic explanation. I then, in section four, suggest that maybe the intentional explanation of behaviour is not undermined by mechanism. This will take the form

of a presentation of Dennett's compatibilism. In the following section I put some pressure on Dennett's position. In the main my concern is simply to cast doubt upon the legitimacy of his compatibilism. I focus primarily on the notion of rationality he exploits and suggest that it does not accord with either our intuitions or our requirements. Finally I show that Dennett's *intentional stance* is better understood as a pseudo-intentional stance and that although it is successful as an option for a kind of explanatory compatibilism the resultant compatibilism is weak and fails to find a place for genuine moral responsibility in a mechanistic world.

2.0 The ascription of responsibility requires the adoption of the intentional stance

I wish to begin by presenting what I take to be the intuitive or common sense orthodoxy regarding the ascription of responsibility. Basically I examine the conditions under which ordinary people would be inclined to ascribe praise or blame for a particular piece of behaviour and when they would be inclined to withdraw such ascriptions. Given that such people generally begin from a position where they do ascribe praise and blame for behaviour we might find that the conditions under which they retract their ascriptions are more instructive.

The judgement of responsibility involves essentially two parties - the judge and the judged.[2] These may be the same person as when I judge myself responsible for something that I did. Most of the time, however, at least two individuals are involved. When responsibility is at issue we are not concerned with the goodness or badness of behaviour. So it is not as if we judge people responsible for their bad actions and not their good ones or vice versa. Rather we judge on the basis of the circumstances of the behaviour. That is, we evaluate the choices open to the person and their reasons for behaving as they did. Making a judgement of causal responsibility stick requires a presupposition that the person is an agent. So we do not assign responsibility to most animals, mechanisms or micro-organisms. This is because we cannot generally make sense of their behaviour from a perspective which would imply that they are causally responsible for their behaviour. If we could understand their behaviour from an internal perspective such as to suggest that they are agents, that is that they are causally responsible for what they do, we might begin to assign responsibility. But as things stand we can only understand, and that generally in an impoverished sense, why they did what they did. We say that it was unfortunate or lucky but not despicable or admirable which would imply agency. From this discussion it is clear that one of the first things we need to address in trying to gain a clear understanding of the conditions under which we assign responsibility is the notion of agency.

When we discover that a person behaved as they did without having any control over the behaviour at any stage we are inclined to withdraw any ascription of praise or blame. Thus,

---

[2]This, and much of the rest of the paragraph, owes much to Thomas Nagel's "prephilosophical" discussion of responsibility on pp. 120-124 in *The View from Nowhere*.

suppose that Josephine suddenly smashes the priceless Ming vase she was holding and we discover that at that moment she was victim to an isolated but massive epileptic discharge which caused a complete yet temporary loss of motor coordination. Under these circumstances we would tend not to blame her in any but the most trivial sense. The trivial sense in which we would blame her is that in which we say that it was Josephine as opposed to another person in whose care the vase came to an end. Such blame is not moral blame and has no impact on the question of responsibility. The reason we would not blame Josephine in a stronger sense, such as the sense we would employ if instead of the epileptic discharge we discovered that the reason for the breaking of the vase lay in the fact that she coveted the vase, knew that she could never own it, and desired that no-one else have the pleasure, is that there is a strong sense in which she did not *do* anything. In fact Josephine had no way of knowing what was going to happen to her and so could not even have known not to handle the vase then. We say that with regard to that instance of behaviour that it was mere bodily movement as opposed to action, that in the breaking of the vase Josephine was no more an agent than the wind would have been if it had blown the vase over.

The question posed by this example is one about the nature of agency. What conditions do we require to be fulfilled in order for a justified ascription of agency? First we need to draw a distinction. Of any behaviour it is either an action or, shall we say, a mere bodily movement. That behaviour which is action is the product of an agent whereas that behaviour which is mere bodily movement is the product of some force beyond the control of the agent. Josephine's behaviour was mere bodily movement in this sense. Depending on what sort of explanation is true of behaviour we would term it action or bodily movement.[3] So there is basically a single, and apparently simple, criterion for agency - that the individual acts. When the individual does not act, that is when we can explain her behaviour in a way that the reason for the behaviour lies beyond the control of the person, she is not an agent. While this rough and ready characterisation of agency is simple, its simplicity is its flaw for the question has

---

[3]To avoid being accused of begging the question against compatibilism I would like to point out that on all accounts, libertarian and otherwise, if an intentional explanation of behaviour is not true then there is no action. So I am not building into my account any idea that behaviour that is mechanistically caused is strictly speaking not action. From the compatibilist perspective it is obvious that such behaviour does count as action if there is also an intentional description thereof.

been insufficiently answered and has merely been removed to a different level of analysis. Specifically it is now incumbent on us to determine what conditions must be fulfilled for action to have taken place.

The common sense notion of action is unfortunately rather impoverished and so my treatment here probably takes the discussion somewhat beyond that of pre-reflective intuition. I suggested above that the question of agency and hence that of action is to be settled by the nature of the explanation of behaviour. In the breaking of the Ming vase I suggested two differing scenarios. In the first we saw Josephine breaking the vase because of an epileptic seizure. Although it would be ponderous to say, what we really saw, given the explanation, was Josephine's body breaking the vase. That is, we saw mere bodily movement. In the second scenario I postulated that we witnessed very similar looking behaviour only the reason for the behaviour was different. In this case I suggested that Josephine a) wanted the vase b) knew she could not have it c) desired that no-one else have the pleasure of owning it, and d) believed that by breaking the vase this desire would be realised. Short for spelling out the reasons for the behaviour in this way we would probably just say that Josephine broke the vase out of spite. Here we see action.

Action occurs if there is a description of the behaviour under which the behaviour is intentional. Intentional descriptions or explanations are simply ones which show the behaviour to be a *doing* as opposed to a *happening* or in Nagel's terminology an *action* as opposed to a *phenomenon*. On the first description of the breaking of the Ming vase we had a *happening* while on the second we had a *doing*. Rendering a piece of behaviour into a *doing* involves explaining it using a particular purposive or goal related terminology. Words like intending, desiring, wanting, choosing, believing, suspecting, and assuming, among others, feature in this terminology. Basically intentional explanation of behaviour picks out the agent as a source of the reasons for the behaviour and describes those reasons in the language of intentionality. The question as to what we mean by the agent being the source of the reasons for the behaviour is answered in the notion of an agent having ends or goals. Given certain ends that the agent has and certain beliefs about how to achieve those ends we are given a rationale for the behaviour. Intentional explanations, in Davidson's terminology, *rationalize* behaviour. (Davidson, (1980) p. 3) Not all explanations containing intentional terms function in this way according to Dennett.(Dennett, (1978) p. 235) Some explanations cite intentional phenomena

as causes of behaviour yet fail to make the behaviour *reasonable*. Consider, my belief that the car would hit the pedestrian caused me to freeze. As it stands this has the appearance of an intentional explanation but is not one. The citing of my belief does not make my behaviour reasonable in this case, in fact the explanation is a purely causal one.[4] On the other hand if we said that I yelled out a warning because of my belief that the car would hit the pedestrian then we would have an intentional explanation of my action. My belief in this case gives a rationale for my behaviour.

The simplest and most widely used example of intentional explanation is where we say that the individual had a belief-desire pair which was the reason for the behaviour. So, for example, I desire to quench my thirst (my end) and I believe that drinking the water will quench it (the means) and so I drink the water. The explanation here is purposive. There is a particular end which I have, namely the quenching of my thirst, which I believe will be satisfied in a particular way, namely by drinking the water. This differs from the non-purposive explanation of Josephine's breaking the Ming vase while having an epileptic seizure or my immobility at the potential accident.

Of course things are not generally as easily characterised as I have presented them. Consider my drinking of the water but imagine it to be a poisonous mixture for killing aphids. Under one description I intentionally drink the mixture but under another I do not. Did I intentionally drink the liquid? Yes. Why? Because I thought it was water, and I desired to quench my thirst and believed that drinking water would do so. Did I intentionally drink the poison? Well, no. Explain? I desired to quench my thirst, believed that drinking water would do this for me and so initiated the action, but, if I had known that it was *poison* I wouldn't have because I have another belief that says that the cost of drinking poison far outweighs the satisfaction of my thirst. So I didn't intentionally drink the poison although I did intentionally drink the liquid that was poisonous.[5] When we have a case like this and we are trying to work

---

[4]I am not intending to present a position here on the issue of whether intentional explanations are causal explanations or not. What the example seeks to show is that there is a clear difference between citing intentional *states* as causes of behaviour and intentional *explanations* which rationalise behaviour by citing such states.

[5]This example is similar to Davidson's example of the action of turning on a light being

(continued...)

out whether a person is blameworthy or not we look beyond the simple intentional explanation to the existence of what can only be called extenuating circumstances. Thus if I was present at the time in which the deadly concoction was prepared and was fully cognizant and was not otherwise engaged then it is implausible for me to claim that I thought it was water and so we might attribute blame to me for the action.

Although it is sometimes the case that we do not attribute praise or blame, that is responsibility, for an action it is clear that when we do attribute responsibility it is only for actions. Which is to say that action is a necessary but not sufficient condition for the attribution of responsibility. We have seen that action occurs when we can successfully adopt an intentional explanation of the behaviour and so an intentional explanation is a necessary condition for responsibility. In no circumstance where there is not an intentional explanation of behaviour do we assign responsibility.

The upshot of this discussion thus far goes something like this. The notion of moral responsibility depends on the idea that at least some behaviour evinced by a person is blameworthy in the sense that the person had some choice in the matter. Such behaviour is typically called intentional behaviour or action. The idea is that moral responsibility is attributable when the behaviour is intentional, that is whenever it involves a bodily movement coupled with a propositional attitude that plays a causal role, for example a motivational belief-desire pair. When there is merely a bodily movement, or even where there is a bodily movement coupled with the relevant propositional attitude but where the propositional attitude does not play a motivational, that is causal, role[6], we resist talk of there having been an action

---

[5](...continued)
identical with alerting a prowler to one's presence. See *Actions, Reasons, and Causes* pp. 4 - 6. There Davidson shows that under the description of turning on the light the behaviour is an intentional action because it is rationalised by a desire to turn on the light and a belief that the movement in question would fulfil this. But under the description of alerting the prowler the behaviour is not an intentional action because the relevant beliefs and desires do not rationalise the behaviour.

[6]This scenario is that of the Epiphenomenalist position on the mental. If an epiphenomenalist wishes to hold onto the notion of responsibility it cannot be in the form of our ordinary conception. This does not mean to say that such a move is impossible it is just beyond the scope of the present work to evaluate it. I am merely seeking to cast doubt on one particular
(continued...)

and we do not assign responsibility to the person. A further condition that might be placed on the attribution of moral responsibility stems from the conviction that the propositional attitude must not only be motivational but must also not be caused by an antecedent physical chain.[7] This is just the idea that if we are going to assign blame for something we should lay it at the feet of the perpetrator. If the propositional attitudes involved in action are causally linked to an antecedent physical chain then the ascription of blame becomes a somewhat arbitrary affair. In fact it begins to look as though the ascription of moral responsibility is just a case of settling on a scapegoat. To preserve the non-arbitrariness of the ascription of moral responsibility we have to have a conception of action that does not allow for passing the buck. This conception is, I submit, that folk conception discussed above.

---

[6](...continued)
position, namely Dennett's.

[7]Once again this might seem to beg the question against compatibilism. In fact it does and I am well aware of this. The fact is however, that I am not attempting to argue against, or for, any particular position here. I am not even presenting an argument. All that I am concerned to do at this stage is to draw out the details of our 'folk' intuitions on these matters. That these intuitions are unpalatable from a compatibilist stance is not something that I can really change.

3.0 Intentional Explanation apparently undermined by Mechanism

In the previous section we saw that our intuitions regarding the ascription of responsibility are that we can only make a legitimate ascription if the behaviour in question is action. We also saw that *action* is characterised by our being able to use a true intentional explanation of the behaviour.[8] In this section I propose to show that a different type of explanation, namely mechanistic explanation, undermines intentional explanations. Once again I present the view from common sense. Where one type of explanation undermines another we have what is called *explanatory incompatibilism*. I wish to suggest that our intuitions tell us that with regard to mechanist versus intentional explanations such incompatibilism is true. If this is correct *and* it proves to be the case that mechanistic explanations are possible for behaviour that we previously used intentional language to explain then our ascriptions of responsibility are in trouble. The reasoning here is simple. If there is an intentional explanation of behaviour then it is action. If there is action then we can be held morally responsible. Mechanism undermines intentional explanations and so there is no action. Because there is no action there can be no ascription of responsibility.

What are mechanism and mechanistic explanation? Simply stated mechanism is the idea that there is a causal chain of antecedent physical events leading up to any particular event. This differs from determinism in that the causal chain in question can be probabilistic in nature as opposed to the deterministic idea that the events which are causally linked are linked universally and without exception. So mechanism allows for the possibility of uncaused causes but they must be physical. The two basic features of mechanism are that it is event-causal[9] in nature and that antecedent *physical* events are always the causes. There are many

---

[8]Once again, from the compatibilist perspective this begs the question. Compatibilists obviously accept that we do not have to use intentional explanations of behaviour for it to be action. What is going on here is that the folk perspective on action and causation involves what we might call 'agent causation' which requires reasons to be the only causes of behaviour that is to be called action. For more on 'Agent Causation' see Sect 5.2, and Taylor, R *Metaphysics*, pp. 51 - 53.

[9] 'Event' causal as opposed to 'object' or 'agent' causal. Event causation is the standard conception of causation in physics. It could be contrasted with the notion of an object in itself being a cause. This is, at this stage, mysterious to me but I suggest the idea as a foil to my use of the term 'event-causal'.

cases of mechanism with which we are familiar. The obvious one is the case of the working of a mechanical object but chemical reactions, quantum mechanical events, and neurological events are all examples too. Mechanistic explanation is explanation that cites physical events as the causes of other events. So, for example, explaining why the car will not start by saying that its spark plugs are dirty and so no spark is igniting the fuel is a mechanistic explanation. We cite an antecedent physical event, namely the plugs' being dirty, to explain another physical event, namely the car's failure to start.[10]

With regard to the main thrust of this project we need to gain an understanding of what sort of mechanistic explanation could be used to explain human behaviour. We then need to ascertain whether or not it does undermine intentional explanation. The most convenient mechanism to use is probably that of neurophysiology. It is granted that at this point we do not have a very sophisticated neurophysiological theory but we must accept that at this point we do not have *any* very sophisticated mechanist theory about human behaviour. I am going to assume that we have such a theory so that we can see whether the explanations that it yields do undermine intentional explanations. The hypothesised theory is such that it is able to explain and predict all human behaviour except that which is caused by forces immediately external to the person.[11] So instances of sneezing, hiccoughing, laughing, moving one's limbs, and twitching one's fingers would all be explained by the theory. The method of explanation would be to show that various changes, both electrical and chemical, in the nervous matter of the body cause specific muscle flexion or extension which in turn are the movements described above. Any behaviour that occurs sans external force is assumed by the theory to originate in this manner and is thus in principle explainable in this way. So we would have complete causal explanations for Josephine breaking the Ming vase, my drinking either the water or the aphid poison, and my yelling out at the scene of the potential accident. Furthermore the explanations would not mention anything like purposes, ends, or goals in their systematic breakdown of the behaviour.

---

[10]While the plugs' being dirty might not be considered an event in the strictest sense, it implies an event, namely spark failure. Basically I employ a very broad conception of what counts as an event here. States, processes, and standing conditions would all count.

[11]This characterisation is basically the same as that suggested by Malcolm in "The Conceivability of Mechanism".

Dennett furnishes us with a useful example.[12] He suggests that we suppose that a man is found who cannot, or will not, say the word "father". In all other respects he is perfectly normal and expresses surprise at his 'inability' to say 'the word I can't say'. Dennett proposes that a psychoanalyst might offer a plausible explanation of this behaviour in terms of unconscious hatred and desires and beliefs the man has about his father. This explanation, it is assumed, makes the man's failure reasonable. In other words it rationalises the behaviour. Such an explanation would be a paradigm case of intentional or purposive explanation. But, we are to further imagine that a neurosurgeon arrives on the scene. He has at his fingertips the vast and complex theory envisaged above. This neurosurgeon establishes that there is a tiny lesion in the speech centre of the man's brain caused by an aneurysm and that lesion is causally responsible for the man's language deficiency. It would seem, intuitively at least, that the intentional explanation has been undermined. A physical event in the brain was the cause and the event cannot be justifiably called the man's reason for not saying the word. The mechanistic explanation renders the behaviour as a *happening* as opposed to a *doing*.

Does a principle emerge from this example that mechanistic explanations undermine intentional explanations? On the face of it the example *is* one where the mechanistic explanation undermines the intentional one. It is very difficult to argue with science, particularly when science is in a position to point to the cause for the behaviour in a way that the non-scientific intentionalist cannot. But is there any reason to suppose that there is a general point to be made here that two different yet complete explanations of a single event cannot both be correct?[13] There is nothing *in the example* that makes such a supposition obvious. Other examples may exist where we are quite happy with both explanations, or alternatively we might have an ontology that explicitly renders both explanations compatible.

One case of an ontology that does this would be reductive materialism. The reductive

---

[12]This example comes from the opening paragraph of Dennett's "Mechanism and Responsibility" in *Brainstorms*.

[13]For an interesting discussion on the possibility of 'explanatory exclusion' see Chapter 13 of Jaegwon Kim's *Supervenience and Mind*. He argues that the notion of explanatory exclusion or incompatibilism is fundamentally right but does not base his position on a single example such as that which we have just examined. For a similar defence of this notion see also Norman Malcolm's "The Conceivability of Mechanism".

materialist believes that mental items are directly reducible to brain states. So an explanation that picks out beliefs and desires as the reasons for behaviour is an explanation that picks out neurological states as the reasons for behaviour. Whether we opt for one or the other is up to us, but generally it will prove to be easier to use the intentional explanation because of its lack of complexity. I am not going to examine this compatibilist move further since it has no bearing on Dennett's work. The fact is, however, that this position usefully demonstrates that there is not an obvious in principle reason to claim that mechanistic explanations undermine intentional explanations. The question we must ask in the face of this is whether it is indeed true that common sense tells us that mechanism does undermine the intentional.

I think a different sort of example is instructive here. It is of a form that Dennett suggests is an intuition pump.[14] The example is that of the behaviour of a wasp called *Sphex*. Dennett cites from Wooldridge:

> When the time comes for egg laying the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyse but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralysed cricket, which has not decayed, having been kept in the wasp equivalent of deep freeze. To the human mind, such an elaborately organised and seemingly purposeful routine conveys a convincing flavour of logic and thoughtfulness - until more details are examined. For example the wasp's routine is to bring the paralysed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If, while the wasp is inside making her preliminary inspection the cricket is moved a few inches away, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again the wasp will move the cricket up to the threshold and re-enter the burrow for a final

---

[14]An intuition pump is simply an example that is used by philosophers to pump the readers' intuitions along the desired route. Dennett believes that we should be constantly aware of this possibility so as to prevent being sent on an emotional roller-coaster to a destination that we would not ordinarily accept. He is, of course, not adverse to exploiting them himself in times of need. See, however, Chapter 1 "Please Don't Feed The Bugbears" of Dennett's book *Elbow Room*, for a discussion.

> check. The wasp never thinks of pulling the cricket straight in. On one occasion, this procedure was repeated forty times, always with the same result.

Sphex's behaviour is shown to be tropistic. That is, it is shown to be completely rigid within the parameters of the wasp's existence. On our first meeting with *Sphex* we are probably likely to patronisingly exclaim at her cleverness. We would even say that she acts, since there is a genuine intentional explanation of her behaviour. That is, there is an explanation of *Sphex's* behaviour that exploits intentional terms and renders the behaviour reasonable. The explanation would be something like the idea that she knows that while she was away something could have taken up residence in her 'deep freeze' and so before storing her food there she checks to see that all is okay. This entails that she has certain beliefs about predators, desires for her off-spring and their food, and beliefs about how to secure those desires. Then the zoologist arrives and demonstrates the rigid mechanism of her behaviour. We rapidly, and with no fear of incoherency, retract our claim that she acts. We in fact say that it is only *as if* she acts, and that it is not that she really believes and desires but only behaves *as if* she had these intentional states. Furthermore we would say, unhesitatingly, that the intentional explanation is wrong.

This example, far from pumping the intuitions, clearly demonstrates the fact that on an intuitive level we do not allow intentional explanations to co-exist with those of mechanism. The reason for this is once again not obviously an in principle reason. I want to suggest that in the majority of cases the reason is a chauvinistic one. We retract the intentional ascription because we do not want to entertain the possibility that we are 'waspish' in our behaviour. For us to entertain such a notion would, it appears, lead us to a position where there is no longer any meaning in what we do. In fact we would no longer really be doing anything - there would just be phenomena in the world. Basically the mechanistic explanation reveals to us that *Sphex* is not able to do otherwise, that she is completely at the mercy of hard-wired 'instincts' and it is a frightening thought that the same could be true of us. This would seem to be a very strong emotive reason for not accepting straight off a compatibilism between the two sorts of explanation. We embrace the explanatory incompatibilism and hold that our behaviour, at least that which we want to call action, is subject only to intentional explanation.

In a way this reason is not at all a good reason for it begs the question against

compatibilism. Basically it starts from an assumption that mechanistic explanation cannot be true of meaningful behaviour. This issue is precisely one which a compatibilist would seek to deny. So if we are to proceed on this matter we have to find a logical reason for supposing that compatibilism is at least not obviously true. That is, we have to find a logical reason to assert that mechanistic explanation cannot be true of action. A classic exposition of such a reason is provided by Norman Malcolm in his essay "The Conceivability of Mechanism".[15] Malcolm in effect suggests that our intuition that the intentional explanation of *Sphex's* behaviour is undermined by the mechanistic explanation is sound. He reasons that the theory exploited by the mechanistic explanation is one which provides sufficient causal reasons for behaviour. So everything done by *Sphex* (or a human for that matter) is completely accounted for by the theory. It follows from this, says Malcolm, that whatever intentions feature in the intentional explanation of the behaviour have nothing to do with the behaviour. The behaviour would have occurred as it did whether the intentions were present or not. When we take account of the *causal completeness* of mechanistic explanation then we see that there is no space left for causal explanation of behaviour in terms of intentional notions. Given that meaningful or purposive behaviour is that behaviour correctly explained by intentional talk and that mechanistic explanations render such talk superfluous there is more than a grain of truth to the common sense assumption that mechanistic explanation cannot be simply true of action.

The basic idea, then, behind our incompatibilist intuition can be summed up as follows. In our ordinary understanding of the notion of action it is the mental states and the logical connection between them that does the causal work in bringing the behaviour about. A mechanistic explanation essentially tells us that this has not happened. Some group of physical events do the causal work and so there is no space left for the mental states to have done it. To claim that there is space would be to say that all actions are in fact over-determined. That is, all actions are caused by twice the number of causes that are sufficient to bring the behaviour about. This offends against our cherished principle of parsimony, namely the general heuristic which states that if two theories are equally tenable the simpler one is to be preferred. Thus we see a *prima facie* case for the incompatibilist intuition.

---

[15]For further discussion of this issue see Jaegwon Kim's *Supervenience and Mind*, Chapter 13 where he echoes Malcolm's concern.

The upshot of this discussion is that mechanism does displace the purposive. Is this enough to allay fears of 'waspishness'? Clearly not, for it may prove to be true of us that our behaviour *is* explainable from a mechanistic stance. In fact, for the purposes of discussion we have assumed such an eventuality and have sought to probe its implications. What we have found is that a chauvinistic prejudice has alerted us to an incompatibility between two sorts of explanation for which we can at least provide some unsophisticated common sense reasoning. The argument does not justify our chauvinism but it does give substance to our retraction of the intentional explanation of mechanistically explained behaviour. I note again that as it stands this argument does not refute a compatibilist position that makes intentional events such as wantings, believings and desirings identical with complex physical events such as particular neural configurations or whatever. Such a position clearly allows for an explanatory compatibilism but it is not clear whether we could reconcile it with our intuitions about agency. Be this as it may, we are concerned here only to provide an illumination of folk-intuitions regarding mechanistic and intentional explanations. The basic fact that I take to have been established is that ordinarily we accept that mechanistic explanation displaces intentional explanations.

4.0 Dennett's Compatibilist Move

In the previous section we saw that there is at least a *prima facie* case for explanatory incompatibilism. We also saw that one possible move that can be made to avoid this incompatibilism is to adopt a version of reductive materialism. I suggested that this might be a successful route to go but that I would not be examining it. I now wish to focus on Dennett's work in this area. He does not opt for the reductive materialist line but still tries to preserve some kind of explanatory compatibilism. Basically I am going to examine the manner in which he attempts to effectively bypass the *prima facie* case previously established. Central to his perspective on this matter is his presentation of what he calls the *intentional stance*. This is ostensibly a stance we can adopt toward certain things in predicting or explaining their behaviour. I shall begin by providing an analysis of this stance and the two mechanistic stances that Dennett distinguishes it from. I then show that on Dennett's account there is a *prima facie* case for accepting an explanatory compatibilism. I draw out the details of how this case can, in Dennett's eyes, be fleshed out toward a full-fledged explanatory compatibilism.

4.1 The Intentional Stance Distinguished and Discussed

Dennett offers us three different systematic methods of predicting and explaining the behaviour of objects or systems. The first two methods that Dennett discusses are mechanistic contenders for the explanation and prediction of behaviour. Dennett introduces them as touchstones for what he calls the *intentional strategy*. The *intentional strategy* is a predictive strategy that is not mechanistic in nature.

The first strategy we can use to explain or predict the behaviour of a system involves adopting what Dennett has called the *design stance*. He uses the example of a chess playing computer:

> If one knows exactly how the computer is designed... one can predict its designed response to any move one makes.... One's prediction will come true provided only that the computer performs as designed - that is, without breakdown. Different varieties of design-stance prediction can be discerned, but all are alike in relying on the notion of a function, which is purpose relative or teleological.... The essential feature of the design stance is that we make predictions solely from knowledge or assumptions about the system's functional design, irrespective of the physical constitution or condition

> of the innards of the particular object. (*Brainstorms* p. 4; for a similar passage see *Brainstorms* p. 237)

To view the design stance in action consider the following example. There is a button mounted on the outer frame of my front door. I predict that when that button is pushed a buzzer will sound. Why? Because when I designed my doorbell I made it so that pushing the button in question would close a circuit. This allows an electric current to pass through an electro-magnet turning it on and causing a metal clapper to move toward it. This causes the clapper to break contact with a metal plate which it needs to be in contact with in order for the circuit to remain closed. As this happens the electro-magnet is turned off, releasing the clapper so that it falls back onto the metal plate allowing the process to begin again as long as the button is depressed. This happens rapidly producing a buzzing sound. Note that these details about the physical operation of the system give me knowledge of the design that I need in order to make the prediction. I am assuming that all of these conditions will be satisfied when I make my prediction. A less complex, and (if you have ever stood outside trying to ring a broken doorbell) less successful, version of the design stance in action using the same apparatus is the functional prediction that some noise will happen when the button is pressed because that is the nature of doorbells.

The second strategy that we can adopt in predicting the behaviour of a system Dennett calls the *physical stance*. The *physical stance* is straight-forwardly mechanistic in nature.

> From this stance our predictions are based on the actual physical state of the particular object, and are worked out by applying whatever knowledge we have of the laws of nature. It is from this stance alone that we can predict the malfunction of systems.... One seldom adopts the physical stance in dealing with a computer just because the number of critical variables in the physical constitution of a computer would overwhelm the most prodigious calculator. Significantly, the physical stance is generally reserved for instances of breakdown, where the condition preventing normal operation is generalized and easily locatable.... Attempting to give a physical account or prediction of the chess-playing computer would be a ... herculean labor, but it would work in principle. (*Brainstorms* pp. 4-5; pp. 237-8)

Returning to my home-made doorbell example; when the buzzer does not work - that is, there is a failure in designed function and my prediction from the design stance fails - I am

forced to adopt the physical stance. Basically, I reason that there is something preventing the mechanism from operating as designed and so look to it's physical state for an explanation. In this case it might be any of a number of things - a loose connection, a disconnected wire, some fluff between the clapper and the plate, dust jamming the hinge on the clapper, or a faulty or disconnected power source. The simplicity of the system makes it comparatively easy to list and test the possible problems but as systems become more complex the variables increase dramatically and it becomes increasingly difficult to adopt the stance successfully.

One should note however that for any purely physical system this stance is a superior stance to adopt as it *guarantees* a correct prediction or explanation of the behaviour of the system as described. The reason for this lies in the fact that for a purely physical system we have to assume what is commonly called the principle of causal closure.[16] This principle states that any physical event that has a cause has a physical cause. In other words it exploits exactly the same idea that mechanism exploits. So when we are faced with a physical system and we want to predict its behaviour, adopting the *physical stance* toward it means that whatever we isolate as the cause of the behaviour will completely explain the behaviour. Given that the explanation is thus necessarily complete it must also be correct. There is no space left for something else to feature in the 'correct' explanation. So in terms of both success and accuracy the *physical stance* cannot be improved upon. As long, that is, as the behaviour being explained or predicted is that of a physical system. It is only complexity and the restricted nature of calculators (human or machine) that often make the stance impractical.

Both of these stances are *mechanistic* contenders for the explanation or prediction of behaviour. In the course of this project I will talk about the *mechanistic stance* (to be contrasted with the *intentional stance*) and I will have Dennett's *physical stance* in mind. This is simply to avoid unnecessary complexity and possible confusion. The fact is noted, however, that both *design* and *physical* explanations are essentially mechanistic in nature. These are to be contrasted with a different type of explanation, namely intentional explanation. The stance in question is appropriately called the *intentional stance* and involves adopting the following strategy:

---

[16] See for example, Jaegwon Kim *Supervenience and Mind* pp. 280-1, where we find that this principle should be accepted in the face of a commitment to naturalism and physicalism.

> This [the *intentional stance*] tends to be most appropriate when the system one is dealing with is too complex to be dealt with effectively from the other stances. In the case of the chess-playing computer one adopts this stance when one tries to predict its response to one's move by figuring out what a good or reasonable response would be, given the information the computer has about the situation. Here one assumes not just the absence of malfunction, but the rationality of design or programming as well....
>
> Whenever one can successfully adopt the Intentional Stance toward an object, I call that object an *Intentional System*. The success of the stance is of course a matter settled pragmatically, without reference to whether or not the object *really* has beliefs, intentions and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably *are* intentional systems, for they are systems whose behaviour can be predicted, and most efficiently predicted, by adopting the intentional stance. (*Brainstorms* pp. 237-8; for a similar but more detailed discussion see *Brainstorms* pp. 5-7)

The *intentional stance* is characterised just by our ability to construct an explanation of a system's behaviour that cites intentional states as the reasons for the behaviour and that makes the behaviour under examination reasonable. Working with Dennett's chess-playing computer then, we can imagine the following scenario. The computer takes my queen. I wish to explain why it did this. I find that if I ascribe to the computer certain chess oriented beliefs and desires then I can provide an explanation of its taking my queen that is reasonable and intentional. So I say that the computer has the desire to win, the belief that my queen poses a major threat to its king, and the belief that having ones own king threatened is an obstacle to victory. Now, assuming rationality, I can say that the reason the computer took my queen lies in those beliefs and desires. Consider the strand of practical reasoning that demonstrates this: If one desires to win, then one ought to remove obstacles to victory. I [the computer] desire to win, so I must remove obstacles to victory. If my opponent's queen poses a major threat to my king, then it is an obstacle to victory. I believe that said queen does pose such a threat, so it is an obstacle to my victory. Putting the two conclusions together we find the behaviour rationally and intentionally explained.

4.2 Dennett's Compatibilist Move

Dennett examines two basic reasons people might have for claiming that mechanistic explanations displace the intentional, or at least that there is an antagonism between the two sorts of explanation. The first of these, suggests Dennett, is that there is an absence of an assumption of rationality at the mechanistic level, while the second pertains to the idea that reasons are distinct from causes.(Dennett, (1978) p. 247) I shall now examine Dennett's treatment of both issues. The idea here is not to engage in a dispute about what Dennett says but simply to gain clarity on how it is that Dennett thinks he can avoid the incompatibilist position. We must note at the outset that his description of the *intentional stance* is such that it seems clear that he can do so. Consider that we say that the explanation of the behaviour of the chess-playing computer from the intentional stance at least has the appearance of being a genuine intentional explanation. It fulfills all the previously established criteria for an explanation to be intentional, that is it cites intentional states as the reasons for the behaviour and it rationalises that behaviour. At the same time however we also know that the computer does what it does merely as a result of certain electrical processes in its central processing unit. We even have Dennett's admission that "attempting to give a physical account or prediction of the chess-playing computer would be a ... herculean labor, but it would work in principle. (Dennett, (1978) pp. 4-5; pp. 237-8) So there does seem to be a case for saying that the two sorts of explanations are *in fact* compatible. Just because the mechanistic explanation is possible does not mean that we retract our intentional explanation. The fact is, however, that this case only establishes that we can use intentional explanations of mechanistically caused behaviour, but just *presenting* the example does not in itself make a case for compatibilism. It simply bolsters the idea that the use of intentional idioms can be seen as a useful anthropomorphisation. The case does not establish the further issue that *intentional* behaviour can be legitimately explained mechanistically. This is what Dennett needs to accomplish in order to develop a full-fledged explanatory compatibilism.

Let's tackle Dennett's first concern, namely that the supposed antagonism between intentional and mechanistic explanations stems from the absence of a presupposition of rationality in the latter case. The basic idea is that when one does not assume rationality the behaviour cannot be seen as action since reasons do not feature in the causal milieu of the behaviour. Dennett seeks to allay this fear by pointing out that "the absence of a

presupposition of rationality is not the same as a presupposition of non-rationality"(Dennett, (1978) p. 243). It is precisely this claim that grounds Dennett's compatibilism. The argument he seeks to refute suggests that behaviour is rational only if the presentation of logically relevant considerations can influence the behaviour. But this means that rational behaviour excludes the possibility of it being the effect of sufficient conditions independent of the agent's deliberation. Mechanistic explanations by definition pick out such independent sufficient conditions and so cannot explain rational behaviour, or action as it is popularly called. The upshot is that intentional explanations have as their domain the truly rational while mechanistic explanations have the rest.[17]

The question we have to answer, and the answer we have to understand, is how mechanistically explained behaviour can indeed still be rational behaviour. Dennett's short answer was simply to assert that failing to pitch our enquiry at the level of reason giving is not to be in a position where reasons cannot be given. The longer answer involves re-commissioning *Sphex*. We saw earlier that *Sphex's* behaviour was tropistic. Dennett suggests that the only reason we retract our intentional explanation of her behaviour lies in the fact that the mechanism of her behaviour is so simple. He says that any *simple* mechanistic explanation of a bit of behaviour will disqualify it for a plausible intentional explanation. But what happens if as wasp designers we try to enlarge *Sphex's* tropistic reactions to the environment so as to create a more rational fit to whatever nature puts in her path. Imagine that we program her to hurriedly retrieve the cricket instead of getting stuck in the hole-checking subroutine and we program her to fly around and seek the obviously external danger. After a while she would behave in a way that we would not even consider tropistic. Explaining what she does *mechanistically* might take volumes of material. Is there any plausibility in claiming that her behaviour is *merely* mechanistic? It is true that it is mechanistic, but what is the force of the *merely* in this case? What would be its force if we were handed *twenty* volumes of fine print? Dennett contends that the only force would lie in the fact that ultimately the organism, human or wasp, is only imperfectly rational.(Dennett, (1978) p. 245) What this means is that our original regard for the rationality of a system stems from an internal perspective but from an objective point of view we find that behaviour just does not always measure up to what the

---

[17]Dennett discusses this in "Mechanism and Responsibility" p. 244. For a more detailed exposition see MacIntyre, A.C. "Determinism" in *Mind*, 1957, pp.248ff.

system ought to do. The implication is, of course, that imperfectly rational the organism may be but rational it is nonetheless.

On Dennett's account we cannot say that mechanistic explanations displace intentional explanations. As his thought experiment regarding the wasp designers is meant to show, it does not follow from a particular mechanistic explanation that the bit of behaviour is or is not rational. Dennett tells us that the fact that a particular response *had* to follow casts no more doubt on its rationality than the fact that the computer's having to answer '25' to the input '5*5=?' casts doubt on the arithmetic correctness of the answer.(Dennett, (1978) p. 246) What we see here is Dennett developing a conception of rationality whereby the rationality of behaviour is simply a matter of the behaviour being logically appropriate to the goals of the system and not the manner in which the behaviour was initiated. Dennett's conception of rationality could be formalised in the following way: Behaviour is rational just in case we can tell a story where there is information that the system could be said to have which logically entails that behaviour.[18] This is a broad conception of rationality and can be contrasted with a narrower understanding. On the narrow alternative, behaviour is rational if and only if there *is* information, in the form of beliefs and desires, that the system has which logically entails the behaviour and where the system either implicitly or explicitly exploits that logical relationship in initiating the behaviour. Dennett's notion of rationality and the alternative suggested is an important issue and is one to which I return in the following chapter. For now I want to simply address the notion of 'information a system could be said to have'.

I established earlier that action is delimited by whether an intentional explanation of the behaviour is true. I also showed that intentional explanations typically cite beliefs and desires as the reasons for behaviour. In the case of the chess-playing computer we saw that certain beliefs and desires could be ascribed to the computer which made its action of taking my queen reasonable. These beliefs and desires are what is meant by 'information a system could be said to have'. Consequently it is important to understand Dennett's position on the nature

---

[18]It has been suggested to me that this formulation is so harsh that I might be setting up a straw man. I acknowledge that this is a strongly instrumentalist conception of rationality, but I also believe that it accurately sums Dennett's position. I shall return to this in detail later. For now it should be noted that I am not suggesting that Dennett's entire position is so radically instrumentalistic and that the focus here is just the notion of rationality.

of beliefs and desires before we can have an accurate and detailed understanding of his compatibilist move.

The issue of the nature of beliefs and desires also bears strongly on the second reason Dennett isolates for the intuition that mechanistic and intentional explanations conflict. This second reason stems from the idea that reasons are distinct from causes. Dennett suggests that the distinction between reason giving and cause giving fosters the idea that reasoning cannot affect a causal chain and thus that mechanistic explanations displace intentional ones. On this point Dennett urges that the intuition is false. While he accepts the idea that one cannot argue with something that is incapable of understanding he totally rejects the insinuation that an argument cannot affect a causal chain. He points out that the presentation of an argument has all sorts of affects on the causal environment. Arguments set air molecules in motion, make ear-drums vibrate, and have affects in the brain of the audience. Reasons can affect the causal path toward behaviour and causes can themselves be directly and causally related to reasons.

Dennett feels that he has secured the case against the incompatibilist intuition. He has shown that the rationality of behaviour is unaffected by whatever constraints it occurs under and that reason giving is compatible with cause giving. This is the basis of what I have called Dennett's compatibilist move. As I suggested earlier, however, we need to understand Dennett's conception of belief before his position can be laid bare. It is to the notion of belief that I turn in the next section.

4.3 Belief and desire according to Dennett

Recall that in the previous section we saw that Dennett's conception of rationality allows him space to deny the incompatibilist intuition that intentional explanations are somehow undermined by the fact that mechanistic explanations of the same behaviour do not assume rationality. The conception of rationality held by Dennett was established to be such that behaviour is rational just in case we can tell a story where there is information that the system could be said to have which logically entails that behaviour. I also suggested that by 'information that the system could be said to have' was meant 'those beliefs and desires that feature in intentional explanation'. To fully understand Dennett's position then we have to understand his conception of beliefs and desires.

With regard to belief and the *intentional stance* Dennett tells us that:

> Lingering doubts about whether the chess-playing computer *really* has beliefs and desires are misplaced; ...the definition of intentional systems I [Dennett] have given does not say that intentional systems *really* have beliefs and desires, but that one can explain and predict their behaviour by *ascribing* beliefs and desires to them, and whether one calls what one ascribes to the computer beliefs or belief-analogues or information complexes or intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascriptions. (Brainstorms p.7)

This suggests that Dennett's position is one where the beliefs ascribed from the *intentional stance* are simply explanatory aides. As such we could say that for Dennett beliefs and desires are instrumental notions. They are to be attributed to the extent that their ascription makes the stance work as a predictive or explanatory tool. What this means is that the existence of beliefs and desires in this context is dependent on the particular interpretive position one adopts toward the system. Such a position is what Dennett, in *The Intentional Stance*, calls radical interpretationism. Dennett claims, however, that his position also encompasses what he calls radical realism.(Dennett, (1987) p.15) This realism is the idea that whether one has a particular belief or not is a perfectly objective internal fact about one that could, in principle, be discerned by physiological examination. So we need to see how Dennett can claim that belief attribution from the intentional stance picks out objective facts about the system and also that wondering whether the system *really* has beliefs is misplaced.

Dennett provides an answer in *The Intentional Stance*. There he claims (Dennett, (1987) p. 29) that if a system is reliably and voluminously predictable from the intentional stance then it really, in the strongest sense of the word, has the beliefs attributed to it. So basically if the intentional strategy works for a system then it really has the beliefs attributed to it for the strategy to work. Adopting the stance involves adopting a particular interpretive position which makes the existence of beliefs and desires a matter of interpretation and having the stance work makes their existence objective in the sense that if they did not really have those beliefs then the stance would not work. This sense of beliefs being objective facts about the system is slightly different from the realism Dennett identifies however, since discerning the existence of these objective beliefs requires adopting an interpretive position. A physiological examination need not, and perhaps will not, reveal them to us.

Why is doubting that a chess-playing computer (an intentional system) *really* has beliefs misplaced? Simply because the concept of belief entailed by the position does not commit one ontologically to beliefs in the way one would have to be committed in order to have the doubts. Dennett's account of belief is such that it does not make an assertion as to the substance or the manner of instantiation of beliefs but only provides an account of what it is to have a belief, namely to be voluminously predictable from the *intentional stance*.

4.4 A Final Word

I believe that the case for Dennett's compatibilism is now complete. I have shown that he begins from a position where there is an abundantly clear *prima facie* case for some sort of compatibilism between intentional and mechanistic explanations. I used this as a spring-board for a deeper investigation into Dennett's ideas. This investigation revealed that Dennett has what looks to be a solid argument against the incompatibilist notion that rational behaviour cannot be explained mechanistically but that it rests on a particular conception of rationality. This conception of rationality, I claimed, was a broad conception which suggested that rationality was a matter of external assessment. If rational agents could plausibly tell a story about the motivating beliefs and desires of the behaver so that it was clear that the beliefs and desires logically entailed that particular bit of behaviour then the behaviour was rational. I fleshed this out with a discussion of Dennett's notion of beliefs and suggested that he sought to retain a "realist-instrumentalist" position in their regard. Finally I also suggested that Dennett's conception of rationality could be contrasted with a narrower conception and it is to that issue which we now turn. I now begin to put some pressure on Dennett's compatibilist position. I wish to show that there is at least an element of doubt as to the defensibility of his position and that given the strong case for explanatory incompatibilism Dennett's position must ultimately remain questionable.

5.0 Two Different Senses of 'Rationality'

In concluding the previous section I suggested that Dennett's conception of rationality could be contrasted with a different conception. In this section I propose to detail that contrast. The line of argument will be to show that Dennett's conception rests on external criteria and is in an important sense artificial and designed. This is to be contrasted with a conception that makes rationality an internal affair. I argue that this differing conception of rationality fits better with our folk intuitions regarding action and agency. In this way I argue that Dennett's position enjoys a less wholesome place with regard to our ordinary conception of intentional explanation than he would have us believe.

I begin by fleshing out the sense in which a computer could be said to be rational. This move is important because it is exactly this sense which Dennett would have us believe is in place when we adopt the intentional strategy for explanation. Once we have a clear understanding of this notion I discuss what seems to be a case where we are not relying on Dennett's conception and thus show that there can be two senses in which our intentional characterisations invoke rationality. In drawing out this dual use of the concept I intend to open the way to seeing an alternative to Dennett's preferred understanding. Once this has been successfully achieved I examine the alternative in some detail, showing that it represents a notion closer to our folk understanding and that it is primary to the notion exploited by Dennett.

5.1 Rationality a matter of design

We are to determine the sense in which a computer could be said to be rational. In other words this section needs to establish exactly what it is that we are saying when we say for example, that the computer is rational in taking my queen. That this exercise will be instructive is perfectly obvious since the rationality of a chess-playing computer is Dennett's paradigm example in his exposition of rationality and the *intentional stance*. So the question for now is simply about what we mean in asserting the computer's rationality.

To make progress on this matter it will be easier if we have a clear example before our minds to work from. Recall then our previous case of the rational computer. The scenario was relatively simple. A computer is designed to play chess and in the process of a game it

captures my queen. Said queen is, from my perspective, putting pressure on the computer's king and is thus a hurdle to the computer's winning the game. Now we have already seen how the move itself can be said to be rational. The move is rational in the sense that it is reasonable or, analogously, follows logically from the desire to win and certain beliefs about the rules of chess. Thus the move is rational in the sense required from the intentional perspective, for if we recall the definition of intentional explanations we find that it is characteristic of them that they render the behaviour reasonable in this sense. This is *not* the sense of rationality under examination. Rather we are to examine the sense of rationality that is exploited in actually adopting the *intentional stance* and which is involved in the process whereby we are to, in Dennett's words, "assume... the rationality of design or programming." (Dennett, (1978) p.238) So the crucial element to our example lies in the computer's being designed to play chess rather than the fact that what it does as a chess player is successfully described in the language of reason.

When we assume the rationality of either a computer or its design what are we assuming? As a first attempt at answering this question one is tempted to say that we are assuming that the computer will respond to relevant input in a manner that will enable us to construct an intentional explanation of its response. This answer is in a sense true because if the computer did not respond in this way the adoption of the *intentional stance* as an explanatory strategy would fail and so we would be forced to retract our assumption of rationality. But as an answer to our question it does not probe deep enough for the assumption of rationality cannot *just* be an assumption of the possibility of a successful intentional explanation. The rationality of the system is what makes the intentional strategy work. So there must be something further about the system other than its behaviour being explained from the *intentional stance* that makes it rational. We need an understanding of the mechanism whereby a system comes to be in a position where the adoption of the intentional strategy becomes viable.

In attempting to determine what it is to assume that a computer is rational we have seen that the answer lies in understanding what it is about the system that enables us to adopt the *intentional stance*. The clue to this answer is to be found in Dennett's suggestion that we are to assume the rationality of design of the computer. It is the design of the system that enables us to successfully adopt the *intentional stance* and so our question must now be as to what it means for design to be rational.

The rationality of design is typically a product of evolution according to Dennett.[19] The first conglomerations of matter that were able to replicate under the right conditions began the process. Things or circumstances that aided survival or replication were 'good' for them, things that did not were 'bad'. As these replicators[20] evolved into organisms with a means to aiding their survival and replication they brought into being a point of view from which things in the world could be viewed as favourable, unfavourable or neutral. This point of view is what Dennett calls the replicators' 'good'. When a system, be it primordial slime, chess-playing computer[21] or human is realising its good it is flourishing. Through evolution and natural selection these systems develop in such a way as to get better and better at realising their good. This development, according to Dennett, is often characterised by the system making certain tradeoffs such as, for example, one between truth and accuracy of reporting for speed and economy. If a system reaches a point where more often than not it is successful in achieving its good it's design might be said to be optimal. That is, its design is such that generally the system will be teleologically successful.

So for Dennett, design is rational if with regard to the manifest interests of the system it generally ensures their realisation. As Dennett says in *The Intentional Stance* "I want to use 'rational' as a general-purpose term of cognitive approval." (Dennett, (1987) p. 97) When we can approve of what the system is doing, or failing to do, in terms of what we understand its interests to be it is rational. This shows that there is indeed more than a grain of truth in our first attempt at answering the question as to what exactly we are assuming when we assume rationality of design. Recall that the suggestion was that one is tempted to say that we are assuming that the computer will respond to relevant input in a manner that will enable us to construct an intentional explanation of its response. We have seen that when we assume

---

[19]See Dennett's *Consciousness Explained* (1991) pp173-182 for a discussion of this.

[20]The term is originally Dawkins' from *The Selfish Gene* (1976) but is appropriated by Dennett in *Consciousness Explained.*

[21]While it may seem perverse to claim that chess-playing computers evolve we can nevertheless make sense of the notion. The computer has an interest in calculating and evaluating moves swiftly and so processing speed is an environmental pressure for it. If it is to improve as a player it will need to develop faster information processing. The programmer plays the role of nature, and conquest by the grand masters that of natural selection.

rationality of design we are assuming that the system is designed in such a way as to generally guarantee success with respect to its range of interests. We are not assuming that, with regard to the computer for example, when we give it *any* information it will behave appropriately but only that given *relevant* information it will *tend* toward an appropriate response.

One thing that is particularly noteworthy of this account is that the *best* response is not required of the system in order to say that it is rational. By 'best' here we can understand those responses that *actually* satisfy the system's interests. The reason why this is not a requirement lies in the idea that the system may, in evolving into a more successful specimen, develop in such a way, to use an example already suggested, as to favour speed over accuracy. Such development may result in periodic false alarms but as a result may have even better survival value. Consider that a person who is quick to run away from seemingly dangerous situations will sometimes flee from the harmless but is less likely to be caught out on the truly perilous occasion than the person who spends time trying to assess the level of threat. Similarly, the chess-playing computer could in principle calculate all the possible moves resulting from the current piece configuration but doing so would result in time penalties and perhaps the forfeiture of the game.

In making the chief criterion for rationality a matter of the system's for the most part realising its interests Dennett makes rationality a matter of *objective* assessment. A computer's rationality is determined by whether it typically behaves in such a way as to further its interests. It does not matter in the least as to how it comes about that the computer behaves in this way. There is no need for the computer to be aware of what it is doing or even for it have any of the so-called intentional states. As a result its rationality is an external matter, it hangs simply on an objective fit between behaviour and interests. Simply by determining what the system is trying[22] to do and thereby developing a conception of its interests coupled with an observation of the success of its behaviour we can judge its rationality. Consider the lowly calculator for example. Its function is to calculate the answers to various mathematical problems. If it is to 'survive' it needs to calculate accurately, and so accuracy in calculation forms the major part of its interests. Observing its behaviour we find that prompted with all

---

[22]Nothing hangs on this use of an intentional term. I would use a word like 'function' but it has too much in the way of philosophical baggage which I could not possibly address here.

sorts of different mathematical problems it does yield the correct answers. The calculator is thus rational. This, then, is Dennett's conception of what we assume when we adopt the *intentional stance* toward it.

Is there any real difference between a *system's* being rational and a *bit of its behaviour* being rational on Dennett's account? This question is motivated by the fact that earlier (section 4.2) I suggested that the question of a bit of behaviour's being rational is a matter of seeing whether we can tell the appropriate intentional story to explain the behaviour. Now we have a situation where the rationality of a system is a matter of whether its behaviour generally realises its interests. These two characterisations are actually very closely linked. Being able to provide an intentional explanation of behaviour means that the behaviour realises the system's interests. The difference between the two situations lies in the fact that a system which is not rational in the sense of generally satisfying its interests may produce an instance of behaviour which does satisfy its interests. So we can have an irrational system happening to behave rationally. By the same token because the demand of rationality is not perfection in terms of realising interests it is possible for a rational system to behave in a manner which defies intentional explanation. In both cases, however, the systems could not behave in the relevant ways as a matter of course. If they were to behave as described regularly and with few lapses we would have to re-evaluate our ascriptions of rationality or irrationality. In the case of the irrational system that regularly behaved rationally we would have to revise its status to that of being rational and in the case of the rational system that behaves irrationally we would have to say that it is in fact irrational as a system. This highlights once again the fact that the criteria for judging the rationality or otherwise of a system lie outside the system with the publicly ascertainable success ratio of its behaviour. Because of this significant feature in Dennett's conception of rationality I am going to label it *method-independent rationality*[23]. The label merely picks out the fact that *how* the system came to behave rationally is not an issue.

5.2 A different sense of 'rational'

Up to this point I have simply drawn out the conception of rationality that Dennett thinks

---

[23]This is as opposed to *method-dependent rationality* which I discuss shortly. Its main feature is that the *how* of the behaviour plays a vital role in the ascription of rationality.

is exploited when adopting the *intentional stance*. I have shown that this conception is characterised by the idea that something is rational if its behaviour is such that it mostly satisfies the interests of the system. We saw that the means by which this is accomplished by the system are not relevant on Dennett's conception. This sense of rationality, I have suggested, is a product of design or evolution and is objective in that it consists in there being a public, open fit between behaviour and interests. This *method-independent rationality* is the means by which Dennett sustains his explanatory compatibilism. A mechanistic system such as a computer is rational in Dennett's sense and by conceiving of rationality in this way we can easily adopt the *intentional stance* toward it. When we assume that the system is rational we are simply assuming that given its interests it will behave so as to satisfy them. When we adopt the *intentional stance* we characterise its interests in terms of certain beliefs and desires which when put together suggest behaving in a manner which, objectively speaking, would satisfy those interests. By assuming that the system is rational in the sense discussed we assume the success of the intentional characterisation. That the system might be a mechanistic one makes no difference.

I now wish to show that Dennett's conception of rationality is not always the conception we employ when adopting the *intentional stance*. The idea is to show that in some instances of providing intentional explanations we use a very different concept. The present question is, then, whether we are in fact assuming *method-independent rationality* when we characterise systems intentionally. The way forward will be to ascertain what conditions prompt us to retract intentional characterisations and our ascriptions of rationality. I propose that in at least one type of case we withdraw our assumption because we are operating with a different conception of rationality.

The first and most obvious case where we retract our assumption of rationality is when it actually fails. There are two basic scenarios that can be characterised in this way. One would be the case of the onset of say, senility or some dementia. The other would be the case of the talking parrot. When we are faced with a person who is becoming demented but who previously was in perfect command of their senses we are forced into the tragic position of suspecting their rationality. Where before their behaviour was such as to be reasonable given their professed interests and beliefs it becomes less and less reasonable. Eventually we retract any ascription of rationality we may once have made. On similar, but lighter, lines imagine

the following scenario. Paul goes to visit an acquaintance who has recently acquired a parrot. As he sees it he says, "Hello Ms Parrot, I'm Paul." If the parrot were to reply, "Hello Paul, I'm Molly," he would think that the parrot is rational.(Paul has little experience of talking parrots!) Then when, a few minutes later, the parrot again proclaims "Hello Paul, I'm Molly," Paul replies, "So you said, are you having a good day?" to which the parrot responds "Hello Paul, I'm Molly," he retracts his ascription, confused.

In both of these cases we see a retraction of the ascription of rationality. The reasons are both very similar, behaviourally both the demented person and the parrot fail a test for agency - their behaviour cannot be explained intentionally. The problem here is that this is exactly the same reason Dennett would cite for their lack of rationality. As possible scenarios for testing Dennett's position they are not adequate to the task. What we have to find, then, is a species of case where we would retract an ascription of rationality but where Dennett would cling to his. If this can be done we will have discovered that we do not in fact operate toward intentional systems on the level of *method-independent rationality*.

One such case has already been discussed, namely that of *Sphex*. When we first encounter this wasp we ascribe rationality to it. When we discover that her behaviour is tropistic we retract our ascription. She is no longer clever but rather just another example of how nature cares for her own. It is no longer the mental states which we ascribe to her and the logical connections between them that do the work in explaining her behaviour, but a rigidly 'hardwired' instinct. From Dennett's perspective *Sphex* is rational. For the most part her behaviour will further her evolutionary interests. It is, after all, a particularly unnatural occurrence for a scientist to attempt to bamboozle her and her tropism was not designed to detect such deception. So in the case of *Sphex* we see Dennett holding to his *method-independent rationality* ascription while the ordinary person retracts their ascription. This serves to show that ordinarily we do not operate toward those things Dennett classifies as intentional systems on the terms he suggests.

Now that we have an example of a case where our folk position differs from Dennett's position let us examine the details of the folk perspective. The question under examination is simply as to what we are saying or doing when we retract our ascription of rationality from a system like *Sphex* which Dennett would claim is rational. What would we want to be

satisfied in order to say that *Sphex* is rational when the tropistic nature of her behaviour seems to render this impossible? One possibility would be that we require it to be possible for a system to deliberate over what it is going to do in order for its behaviour to be rational. This is not to say that the system *has* to deliberate over what it is going to do, just that it must be *able* to deliberate over what it is going to do.

There are five basic elements to the notion of deliberation. The first of these is that a system can only deliberate about its own behaviour and not about the behaviour of another system. I can, for example, deliberate over what exactly I am going to write next, but I cannot deliberate over whether my reader will understand it or not. I can speculate that my reader will understand it or I can surmise this, but I cannot deliberate over it. The second obvious element to deliberation is that it can pertain only to future behaviour. What I have done or what I am busy doing is not open to deliberation. It is open to recall or cognisance but there can be no weighing up of alternatives. Either the behaviour is done or it is being enacted and so the facts are already fixed. A third consideration which is related to this is that a system cannot deliberate over behaviour that it already knows it is going to do. Because it already knows what it is going to do the facts are once again fixed and no deliberation is possible. Say for example that I have decided to include five points pertaining to deliberation in my discussion of the folk conception of rationality. I cannot now deliberate about whether to include them or not. If I did start to have second thoughts about a particular point I might begin to deliberate over whether to include all five points or not but then I could not say that I already know what I am going to do. Fourthly a system can only deliberate if it believes that in some sense it is up to it as to what it is going to do. So if we imagine my going to a dinner party at a friend's house I cannot deliberate over whether I shall have chicken or red meat since it is not up to me which of these will be prepared. Only if I believe that it is up to me, as I might if my friend had asked me to approve the menu which listed chicken or beef as a third course option, could I deliberate over which I am going to choose. Finally deliberation involves trying to see the logical implications of one's beliefs, desires, and proposed courses of action. Doing this is what must ultimately enable us to make up our minds.[24]

---

[24] Much of this paragraph is prompted by Richard Taylor's section on 'deliberation' in *Metaphysics* pp. 39-40. He of course is not concerned with rationality as such but rather with the (continued...)

Now on the face of it, it may seem clear that *Sphex* could deliberate over what to do when she finds that the cricket has been moved. There does not seem to be any reason to suppose that she does not believe that what she does is up to her, even though we know otherwise. Just as I can deliberate over what to order for dinner even though some of the possibilities I entertain are not true possibilities because of a temporary unavailability, so *Sphex* could deliberate about whether to check the burrow again or to just haul the cricket in even though her circumstances are such as to render the latter an impossible option. Such, anyway, is the case the first time we see her go through the 'hole-checking subroutine'. But after we witness the behaviour many times the description of what is going on must surely change. It must be clear to *Sphex*, for instance, that the danger lies outside the hole since it is the cricket that moves and not, say, an unknown scent emanating from the mouth of her burrow. With this knowledge *Sphex* ought to try to get the cricket inside and assuming that she does try she will quickly discover that she is stuck in her 'hole-checking subroutine'. This would shake her belief that it is up to her what she is going to do and so render deliberation impossible. Assuming that this does not happen however it is also clear that after going through the subroutine for the umpteenth time *Sphex* has more than just a good idea as to what she is going to do when she finds that the cricket has been moved again.[25] In effect then the sheer repetition of experience places deliberation far outside the reach of *Sphex*. Given that she is not able to deliberate over her behaviour she cannot be rational in the folk sense.

A related consideration involved in whether we deem behaviour rational or not is that we require the behaviour to be a result of the agent's proceeding as she does *because* of her awareness of the course of action implied by her interests. So not only must she be able to weigh up alternatives and deliberate about which she will do, she must also do what she does because of this deliberation. In *Sphex's* case her behaviour is a result of certain antecedent events. Any awareness she might have of the course of action implied by her interests can

---

[24](...continued)
problem of free will.

[25]I am of course assuming that *Sphex* has not read Hume's *Enquiry Concerning Human Understanding* and had her faith in inductive reasoning shattered. Whatever the case, however, even Hume accepts the use made of 'conclusions based on experience', he just cautions us to be aware that they are not founded on reasoning per se.

surely play no role in her behaving. For the sake of argument let us assume that *Sphex* is a conscious or an aware system. So we assume that she is cognisant of the course of action implied by her interests. It is clear that she cannot possibly believe that there is something in her hole *because she was in it when the cricket was moved.* So if her checking the hole was to get information regarding the safety situation in the hole, she has it. The movement of the cricket would imply that there is in fact some threat outside the hole. So her beliefs and desires are such as to imply that she ought to get the cricket inside as soon as possible, not that she ought to check the hole for lurking predators. Now we have assumed that she is aware of this and in light of this assumption we can make no sense of her subsequent behaviour being as a result of her awareness. Her subsequent behaviour just does not bear out such a consideration for she does not behave in accord with the reasonable dictates of her awareness. The only way we could make sense of *Sphex's* behaviour in this way would be to add an *ad hoc* desire to behave in a manner contrary to that dictated by reason. This would mean that in being aware of what reason says she should do and in having a desire not to do the reasonable thing, the appearance of *Sphex's* being stuck in the 'hole-checking subroutine' can be rendered understandable. But considerations of simplicity and economy of explanation make this route completely implausible. The mechanistic explanation of her behaviour just rings truer and explains more of her behaviour and so using Ockham's famous razor we must cut away the unnecessarily complex theory which ascribes consciousness and *ad hoc* intentional states.

The scenario we have been considering is perhaps better understood by means of a different example. Recall that we assumed that *Sphex* is aware of what she ought to do. I suggested that on a folk conception we would require such an awareness *and* that the awareness play a role in causing the behaviour. We have seen that for *Sphex* this second condition cannot plausibly be said to have been met. But what is it like for *Sphex*? Why would we say that this sort of scenario is to be excluded from the group of cases we would characterise as rational? The situation is one where the awareness of what ought to be done is *epiphenomenal* in the strictest and most pejorative sense. A little thought experiment will bring the issue out nicely. Imagine that you are *Sphex*. Know that you are aware that you went into the hole to check it out and that while you were there your cricket was moved. You realise that all things considered you had best get the cricket inside quickly because *something is out there*. With this awareness in mind, however, you find your legs inexorably moving you

towards the burrow, cricket once again abandoned just outside. You might mentally kick yourself, saying that this is just against all your better judgements. But when the process happens again and again mild dissatisfaction will quickly give way to alarm as you realise that your body is out of control. Were someone to presume to tell you that your behaviour was rational you might giggle hysterically and say that it was a nice gesture but that your interlocutor needs to face the facts. The discovery that you are the flesh and bone equivalent of a thinking marionette would be terrifying and would undoubtably furnish you with a strong conviction that in all but the most tenuous sense your behaviour is *not* rational. It is abundantly clear then that we require not only an awareness of the course of action implied by the system's interests but for that awareness to play some sort of causal role in the behaviour. On these grounds then, we retract our ascription of rationality to *Sphex*.[26]

What we find emerging from this discussion is a conception of rationality that is based on the *way* behaviour comes about as opposed to the success of the behaviour. The criteria for judging rationality are no longer the objective, publicly ascertainable fit between the behaviour and the ascribed intentional states of the system and lie rather with the process of the system's coming to behave as it does. Of course the notion of the agent somehow causing the behaviour because of the logical dictates of her interests is a mysterious one. Our scientific understanding of causation relies on antecedent *events* being causes and now we have a situation where an *object*, namely the agent, is the cause. Mystery aside, however, this is the understanding that supports and is supported by the folk conception of rationality. Because the *manner* in which the behaviour comes about is so important on this conception I call it *method-dependent rationality*.

The chief characteristic of *method-dependent rationality* lies in its subjective nature. *Method-dependent rationality* is subjective in the sense that it is internal, it pertains to the individual consciousness and is not as such, publicly visible. I do not want to suggest that in being subjective *method-dependent rationality* is merely a matter of interpretive stance or that

---

[26]It has been suggested to me that maybe all that is going wrong with *Sphex* is that she has a very short memory. So she simply doesn't remember that she has already checked her hole. This move cannot work. *Sphex* remembers that she dug a hole, she remembers where it is while she goes off to hunt for a cricket, and she successfully hunts - things that would be completely impossible were she unable to remember from moment to moment!

it is an imaginary or distorted concept. With regard to its existence then, *method-dependent rationality* is to be considered a perfectly objective concept. Due to its internality however it is notoriously difficult to say with certainty that a system is employing *method-dependent rationality* in a particular instance. Consequently I am not going to attempt to provide a systematic breakdown of how we can identify systems that are in fact *method-dependent rational*. Instead I am going to provide a negative method. The idea is to supply a perfectly objective test for whether this form of rationality is even possible for a system. In this manner we can exclude certain systems from the potential group of *method-dependent rational* systems. In other words, if a system passes the test then it *might* be rational in this sense. If on the other hand the system cannot pass the test in that any of the criteria cannot be met due to the very nature of the system then we can rule out the possibility of it being *method-dependent rational* altogether.

There are three basic criteria which we can use. Firstly we require it to be possible to deliberate about the course of action. This is what we might call the *freedom* criterion. Secondly we require that the behaviour follows logically from the system's intentional states, that is, given the system's beliefs and desires a train of practical reasoning can be developed with the behaviour as its conclusion. This would be a *success* criterion and is basically the criterion that Dennett employs for the evaluation of instances of behaviour. All that it requires is that we be *able* to give an intentional explanation. Finally we require that the behaviour come about because of the system's awareness of the logical implications of its beliefs and desires. For want of a better term I think of this as the *causal* criterion. I will briefly discuss each in turn.

The *freedom* criterion first then. Recall that the notion of *method-dependent rationality* pertains to the process of the system's coming to behave as it does. One requirement we have regarding this process is that it must involve the possibility of deliberation. Now deliberation is only possible with an attendant belief in one's freedom, that is that it is up to the system as to what it is going to do. So if a system believed that it had some choice over what it was going to do and that whatever was opted for was in fact up to it, then it would be possible for it to deliberate. There are two basic ways in which a system could fail this criterion for *method-dependent rationality*. In the first place it could simply be the case that the system does not have the required belief. This might come about when the system is either aware of

some constraint on its behaviour or where the system simply believes albeit mistakenly that there is such a constraint. An example of the first sort would be where a person cannot deliberate as to whether she is going to remain on the ground or gently float upwards into the sky. In this case she would be aware that it is not up to her and that her weight plus gravity constitute a physical restraint on her gently drifting into the air. In this respect then she would not believe in her freedom and so her staying on the ground would not be a rational matter. An example of the second sort would be where a person could not deliberate as to whether to switch a light on because she is under the mistaken impression that the power is down. In other words, she believes mistakenly that the future is fixed, she cannot deliberate about turning on the light because she believes that even if she tries to she will be thwarted. The second basic way in which a system could fail this criterion for *method-dependent rationality* would be where the experience of the system was such as to challenge its belief. The obvious example here is the case where there is a breakdown between what the system wills and what it actually does. So I will myself to yank open the door and find that I cannot do so because the door is locked. This would challenge my belief that I am free to leave the room and so in my subsequent behaviour of 'staying put' I could not be considered *method-dependent rational*.

The second criterion that we can use to test a system for possibly being *method-dependent rational* is the *success* criterion. This criterion is simply that the behaviour of the system follows logically from its beliefs and desires. I have already intimated that this is basically the criterion that Dennett employs for the evaluation of behaviour. With regard to systems as a whole however he has a weakened version of the criterion. As we have already seen, he requires only that the system behave in this logically coherent manner for the most part. I think that Dennett has a valuable insight on this matter since making the requirement such that all behaviour must follow logically ultimately excludes all systems we know from being *method-dependent rational*. Dennett has suggested that this is the route some people might opt for in attempting to distinguish between persons and mere mechanisms and he labels it the 'Einstein-Shakespeare Gambit'. As he points out, nothing short of perfection will do and for every supposedly perfect case, thousands of lapses and foibles can be cited. I agree with Dennett on this matter and so remain content with the slightly weaker notion. Systems that fail this criterion are thus obviously ones which Dennett would exclude from *his* group of

intentional systems also. The obvious example is of course that of the lunatic, whose behaviour typically defies reason.[27]

Thirdly and finally, we can employ what I have called the *causal* criterion. I have stated that this criterion involves the system's behaviour coming about because of its awareness of what its beliefs and desires imply that it ought to do. In effect this means that the behaviour does not just come about because of some antecedent event but instead because the system *brings* it about.[28] With regard to *method-dependent rationality* then, the *causal* criterion establishes the type of cause behaviour may have. Events that are antecedent sufficient conditions may not be the cause of the behaviour if it is to be considered rational. The system itself must somehow originate the behaviour, in the words I used earlier, the behaviour must be a *doing* and not a *happening*. Systems whose behaviour is explained by antecedent events that are sufficient to bring about the behaviour in question would thus fail this criterion for *method-dependent rationality*. A perfect example would be the case of *Sphex* discussed above. Due to the very nature of the mechanism behind her behaviour it is impossible for her to be rational in this sense. Because of the tropistic nature of her behaviour it is impossible that her behaviour comes about because of her awareness of the logical implications of her interests. Thus in some cases *facts* about the system enable us to reach a conclusion as to whether the system is even capable of *method-dependent rationality*.

When a system cannot be shown to fail any of the three criteria for *method-dependent rationality* we assume that it is rational in this sense. This is just a matter of charity[29] on our

---

[27]Note that my agreement with Dennett on this matter pertains only to the evaluation of *behaviour*. I don't want to imply that there can be any *action* that is not *method-dependent rational*.

[28]In the present work I am not going to go into the metaphysics of how this is possible, time and space do not allow for it. The important thing is that I am merely drawing out the details of the folk position on this matter. A number of people have addressed the matter, however, see for example Chisholm, R "Freedom and Action" and Clarke, R "Toward A Creditable Agent-Causal Account of Free Will".

[29]The idea comes from Davidson's Principle of Charity. He suggests that we need to credit people with certain things in order to make any sense of what they do. Davidson, obviously, uses the term in a different context to me. See *Essays on Actions and Events*, p. 221 for details.

part and stems from the fact that for the most part we are unable to objectively assess the rationality of behaviour or the system. If a system's behaviour is such that it realises what we perceive its interests to be and there is no obvious constraint on its freedom we assume that it is rational. When systems behave in a manner that violates what we understand to be the logical course of action given what we assume are their interests our assumption of their rationality is called into question. In such a case we rely on the system's introspective testimony as to its motives to see whether there is a logical reason for the behaviour. If there is such a reason we are content with the ascription of rationality and if not we retract it altogether.

Whether this conception of rationality is viable or not in the sense of whether any system is ever able to deliberate and 'move' itself, is beyond the scope of this investigation. It might well be that determinism or mechanism are true of all behavioural systems. This does not alter the fact that from a folk perspective we think that some systems are capable of *method-dependent rationality*. In particular we think that persons are such systems. We may be under an illusion on this matter but that does not affect the conclusion that *must* be drawn from this discussion. In some cases we do assume a different sort of rationality to that proposed by Dennett. His is the broad *method-independent rationality* while the folk conception tends to be the narrower *method-dependent rationality*. This would suggest that there are in fact two different types of *intentional stance*.

5.3 In Summary

What we have seen in this chapter then, is that Dennett's conception of rationality depends ultimately only on a system's behaviour generally satisfying its interests. Using the example of a chess-playing computer I developed a clear model of the conception of rationality which Dennett would have us believe is at the heart of our adoption of the *intentional stance*. I called this conception *method-independent rationality*. I then challenged the idea that this is indeed the conception of rationality we rely on when making intentional characterisations of systems and their behaviour. The method was to find a case where Dennett would have to retain his assumption of rationality and treat the system as an intentional system. but where from an ordinary intuitive perspective we would retract both our assumption of rationality and the intentional characterisation. *Sphex* proved to be a perfect

example. I then developed the details of this alternative conception of rationality and called it *method-dependent rationality*. The single most important difference between the two conceptions of rationality discussed was shown to lie in the fact that *method-independent rationality* is based entirely on the success of behaviour while *method-dependent rationality* is based on the *way* behaviour comes about. Finally I suggested that it is clear that our intuitive conception does differ from Dennett's and that as a result there is a very real sense in which we can say that there are two different types of *intentional stance*. It is to this issue which we now turn.

6.0 Two Intentional Stances

In the previous section I showed that the intuitive conception of rationality, which I have called *method-dependent rationality*, differs substantially from Dennett's *method-independent rationality*. We further saw that *method-dependent rationality* is exploited in some instances of adopting the intentional perspective toward a system which means that Dennett's assertion that we assume what I have called *method-independent rationality* is not strictly speaking true. What is true is that if we adopt the intentional perspective *as proposed by Dennett* then we obviously must employ his conception of rationality. But it is clear that in our normal everyday adoption of this perspective we are in fact employing the stricter notion of *method-dependent rationality*. I suggested that this state of affairs is such that we must conclude that there are two *intentional stances*, Dennett's and the common sensical or intuitive one. In this chapter I am going to examine the two. The basic idea is to show that there is a clear difference between the two types of *intentional stance*. Once this has been achieved it will be apparent that although Dennett has made a case for the compatibilism of mechanism and one sort of intentional explanation, he has not successfully reconciled the notions of mechanism and responsibility. This is because our judgements of responsibility are made from a significantly different position to Dennett's *intentional stance*.

I begin by detailing what for now I am going to call the *folk stance*. This is just the version of the *intentional stance* that relies on *method-dependent rationality*. I show that aside from exploiting a different sense of rationality to Dennett's *intentional stance*, it also demands a strong realism about intentional states. I then turn to Dennett's *intentional stance* and suggest that its nature is such that it is better understood as a *pseudo*-intentional stance. In light of this we turn once again to the issue of compatibilism. I show that Dennett *has* secured the case for explanatory compatibilism, but only with regard to mechanistic explanations and intentional explanations generated from his *pseudo*-intentional stance. He has not done the same for mechanistic explanations and intentional explanations from the *folk stance*. I go on to show that in fact the *folk stance* must remain beyond the scope of Dennett's compatibilist move. Finally I suggest that it is only from the *folk stance* that we judge systems to be responsible for their behaviour. Given this outcome I conclude that Dennett's attempt to secure explanatory compatibilism between intentional explanations and mechanistic explanations is not able to make the additional case for the compatibilism of mechanism and

responsibility. This is mainly because Dennett's attempt aims at the wrong kind of 'intentional' explanation.

6.1 The Folk Stance

We have seen that when we normally adopt the intentional perspective towards behavioural systems a particular conception of rationality is involved which differs from Dennett's conception. This conception of rationality is such that the way in which behaviour comes about plays a crucial role and has been called *method-dependent rationality*. In setting out the ideas that lie behind this conception of rationality we saw that there are three basic criteria that we would want to be satisfied in order to say that the system is in fact rational. Firstly there is the *freedom* criterion, which requires that it be possible for the system to deliberate about what it is going to do. Secondly there is the *success* criterion which is such that we require behaviour to generally fulfil the system's interests.[30] Thirdly there is the *causal* criterion that requires that the system not only be aware of the logical dictates of its interests but also for the system itself to somehow bring about its behaviour because of this awareness. I now wish to investigate what implications, if any, can be drawn from the notion of *method-dependent rationality* for the ontological status of intentional states.

Working with the first criterion it is clear that for a system to be *method-dependent rational* it must actually believe that it is up to it as to what it is going to do. That is, it must believe that it is in some sense free to do as it wishes. Clearly this belief cannot be merely a matter of interpretation. In order for a system to deliberate at all it must *really* believe that what it does is up to it. That is, there must be some mental event within the system that is a belief in its freedom. Merely adopting a perspective where we *interpret* the system's behaviour as involving deliberation will obviously not suffice. This is because an assumption of the presence of the belief is simply not enough to enable the system to deliberate. I don't want to suggest, however, that this belief must actually be a conscious belief at the time of deliberation. In other words I am not suggesting that the system has to be literally aware of its belief in its freedom while it deliberates. On the contrary, I suspect that the belief in question

---

[30]Note that this criterion pertains to the broad notion of behaviour. In the narrower case of action the criterion is much stronger and requires that the behaviour always satisfies a system's interests. See sect 5.2.

is typically unconscious. But we must note that in the very process of engaging in deliberation a system must at the very least tacitly believe that it is free. This would seem to indicate that from the folk perspective we require a certain realism about belief.

The idea that the *folk stance* involves a realist perspective of belief, and indeed all intentional states, is further borne out by the *causal* consideration which features in *method-dependent rationality*. We have seen that this criterion requires that a system's interests, that is its beliefs and desires, feature in the causal story of its behaviour. Which is to say that a system must be aware of the logical implications of its intentional states and its awareness thereof must play a role in bringing about the behaviour. This would suggest that the intentional states mentioned in the explanation of the behaviour must really exist as events that the system itself recognises if we are to say that the system in question is *method-dependent rational*. It would not be enough, therefore, to make the existence of intentional states simply a matter of theoretical interpretation. Making intentional states such as beliefs and desires merely a matter of adopting a particular interpretive stance cannot give them the causal power required. The upshot of this discussion is that in effect the *causal* criterion makes it such that not only are we *able* to provide an intentional explanation of the system's behaviour but that we *have to* provide such an explanation. This is obvious once we realise that the intentional elements of the explanation are events that, by definition, must feature in the causal history of the behaviour. If they do not feature in the causal history of the behaviour then the system would fail the *causal* criterion for *method-dependent rationality* and would not be counted as an intentional system.

What we find developing from this discussion is a conception of the *intentional stance* that is based on typical folk attitudes towards the explanation of behaviour. The stance, which we can call the *folk stance*, is similar to Dennett's *intentional stance* in that it involves the explanation of behaviour based on intentional states and rationality. We have seen though, that the conception of rationality involved is markedly different from that exploited in Dennett's position. *Method-dependent rationality*, as I have called it, focuses on the way in which behaviour comes about. Because of this a large number of behavioural systems that would count as being *intentional systems* on Dennett's conception are excluded from the group of systems so classified on the *folk stance*. Examples would be some of those we have already addressed, such as computers, calculators, *Sphex*, thermostats, and the insane. The other

salient aspect to the *folk stance* is that it demands a strong realism about intentional states. I have already briefly discussed Dennett's position on the ontology of intentional states. I showed that Dennett conceives of his position as lying between strong realism and strong interpretationism. What I seek to establish now is whether Dennett's position is sufficiently realist to be considered compatible with the folk conception outlined above. If it can be shown that his position is not really compatible with the folk conception of intentional states then it will be clear that despite their superficial resemblance, the two different *intentional stances* are totally distinct positions.

6.2 Is Dennett's Conception of Intentional States Realist Enough?

I argued above that if a system is to be classified as an *intentional system* from the folk perspective it must really have intentional states in the sense that they must feature in the explanation of the system's behaviour. I called this a strong realist position. What I seek to show is that there is an unbridgeable gap between Dennett's *intentional stance* and the *folk stance*. So the question that needs to be addressed now is whether Dennett's position on the existence of intentional states is similar enough to the folk position as to make no difference which conception is employed. I hope to show that Dennett's conception is *not* up to the task. This will mean that aside from the difference in conceptions of rationality already established, the two stances also differ radically on the matter of the existential status of intentional states. If this can be established then it will be clear that the divide between the two stances is complete, since both of the crucial elements to the warring explanatory strategies will have been shown to be irreconcilable.

As I indicated above, the crucial element to the realism of intentional states from the *folk stance* lies in the fact that we *have to* advert to them in explaining the system's behaviour. In effect this means that for systems that count as *intentional* systems on the *folk stance*, talk of their beliefs and desires is strictly ineliminable.[31] In comparing Dennett's conception with this conception then, the obvious route is to ascertain whether his position does allow for the

---

[31]By ineliminable I don't, of course, mean that we could not stop using the intentional terms. It is clear, for instance, that on an Identity Theory one would be able to eliminate use of the *words* for beliefs and desires but the beliefs and desires themselves, being brain states, could not be eliminated. This is the sense in which I mean that they cannot be eliminated.

elimination of intentional talk. In other words, we need to see whether Dennett's conception of the existential status of beliefs and desires is meaningfully realist or whether the conception ultimately collapses into mere dependence on the adoption of a particular interpretive stance. Dennett, of course, would want to resist the notion that his commitment to intentional states is purely instrumental but if that is what his position finally proves to be it will just be an unpalatable fact that he has to accept.[32] The idea that intentional talk is basically eliminable on Dennett's conception is not new. It has been raised by Nozick and is addressed by Dennett in *The Intentional Stance*.

Dennett's view, as we have seen, is simply that if a system is reliably and voluminously predictable from the intentional stance (an interpretative position) then it really, in the strongest sense of the word, has the beliefs attributed to it (Dennett, (1987) p.29). This clearly makes Dennett an instrumentalist about belief (Stich, (1990) p.170). He is an instrumentalist about belief because he attributes belief to the extent that it makes the intentional stance work. As an instrumentalist about belief we easily see the interpretive element of Dennett's notion of belief. Beliefs are discernable in agents' observable behaviour when we choose to interpret that behaviour from the *intentional stance*. Now the question we need to resolve is whether this position is, in the final analysis, such that it involves no robust objectivity or realism with regard to intentional states. In other words, whether Dennett's conception commits us to the existence of intentional states as real events or not. With this in mind let us consider an example.

Imagine that a Martian species were to arrive on earth. These creatures are incredibly good physicists, so good in fact, that they can explain our behaviour in its entirety without resorting to the *intentional stance*. They operate toward us on a purely physical, that is, mechanistic, level. Where we are faced with two people moving chess pieces on a chequered board for instance, they perceive a mass of atomic particles behaving in strict accord with the laws of physics. Utilising these rules they are able to say many moves in advance, in fact before the game begins, exactly who will win, and how that person will do so. With regard to our behaviour they have absolute success in predicting what we will do and in explaining

---

[32]For Dennett's thoughts on this matter see his troubled statements to this effect in *Dennett and his Critics*, p. 210.

why we did what we did. What the example presumes to show is that the use of the *intentional stance* is simply a matter of interpretation; if Martians can abandon it with absolute success then the intentional idiom does not occupy a special objective place in the world. Of course the example assumes that we are in fact mechanistic in nature. On this assumption we would obviously be excluded from the group of *intentional systems* as delineated from the *folk stance*. But whether it is true that we are mechanisms of this sort or not the point is still made - when a mechanistic explanation is available we do not have to retain the intentional perspective.

Dennett has a response to this objection and it provides us with the essential clue as to his 'realism' about the intentional idioms. Dennett suggests that if these Martians engaged in a predicting contest of human behaviour with a human they would be amazed at the success of the human, in fact the humans performance would seem magical to them. Why? Simply because that by operating at the physical level they miss a *perfectly objective pattern* - that picked out by the intentional stance. This is meant to show us that the beliefs assumed in the *intentional stance* are objective, real things (Dennett, (1987) pp.25-28). But does it? What is the ontological status of patterns or their elements? For our purposes it is critical to uncover the essence of Dennett's claim that the intentional pattern and its attendant elements (eg. beliefs, desires, rationality etc.) are perfectly objective. It seems that Dennett is saying that intentional states are real because where there are intelligent beings the relevant behavioural patterns are there to be described whether we care to see them or not. But what kind of existence can we grant to a pattern?

To answer a question such as that posed it would be wise to first establish exactly what we are talking about. What, then, is a pattern? This complex topic has been tackled by Dennett in his influential and ground-breaking article, "Real Patterns". [33]According to Dennett a pattern is a candidate for pattern recognition. But this has the air of tautology about it. When is something a candidate of the aforementioned sort? The standard answer appears to be that something is a candidate for pattern recognition if there is some way to transmit all the information needed to replicate that thing without having to transmit the *bitmap*, that is without having to transmit the information bit by bit. In the jargon of pattern-talk we say that

---

[33]See *Journal of Philosophy*, (1991) 88, pp. 27-51.

a pattern exists if there is an *algorithm* which describes it. An algorithm is simply a formula that compresses the information. So, for example, the set of even numbers from zero to infinity is a pattern. Why? Because we can transmit all the information needed to replicate the set (which consists of an infinite number of separate bits of information) in a mere nine words, that is "the set of even numbers from zero to infinity". The phrase "the set of even numbers from zero to infinity" is one algorithm that describes the pattern expressed by 0, 2, 4, 6, 8, 10, 12, 14, 16, 18 etc. Now a pattern need not be as neatly expressible as this in order for it to exist, all that is required is that it be possible to transmit all the information using less bits than the pattern itself occupies. So the number string "0, 1, 2, 4, 6, 7, 8, ... infinity" is also a pattern only its algorithm is "the set of even numbers from zero to infinity including 1 and 7". The two odd numbers in this example are called *noise*. If the noise gets so dense as to render it impossible to transmit the required information in less than the bitmap then the pattern ceases to exist.

To illustrate his position with regard to intentional states Dennett draws an analogy with John Horton Conway's Game of Life (Dennett, (1991b) p.37 ff). Life is played on a theoretically infinite two dimensional grid. The grid divides the area into cells which at any point in time might be alive or dead. Live cells are filled, while dead ones are left blank. Each cell has eight neighbours. Time advances in ticks in the game. The population of the Life world changes from moment to moment according to the following basic rule: Each cell determines the state of its neighbours. If only two neighbours are alive that cell will remain in its current state for the next time slice. If three neighbours are alive the cell will also be alive in the next instant. In all other cases the cell will be dead for the next instant. Now, as Dennett points out, if we apply this rule scrupulously we can predict with perfect accuracy exactly what is going to happen on Life in the next instant, and the instant after that, and the one after that... and so on. By applying the rule we latch onto the physics of Life and are able to utilise an algorithm for prediction. This is analogous to adopting the *physical stance*. The example shows us that where mechanism is true the physical stance yields perfect predictions. Taking a step back from the trees as it were we soon discover that sometimes the cells form configurations that are interesting. Some configurations swim across the grid, others eat any configuration that they come into contact with, and yet others give birth to various swimming configurations. Naturally these come to be named and we suddenly find Life populated by

gliders, *ships*, *eaters*, *glider-guns*, and *puffers*. (See figures overleaf) These 'entities' behave in a consistent way. They persist through time. Where before there was only a grid of either alive or dead cells there is now a veritable zoo of recognisable entities. Viewing Life from this level is analogous to adopting the *intentional stance*, or so Dennett wishes to suggest.

The question Dennett would have us answer is whether there *really* are gliders or whether there are just cells forming patterns that behave *as if* they were gliders (Dennett, (1991b) p.39).It is clear that in the example of Life there are just two different 'translation manuals' for what is going on. One manual tells us about individual cells - it generates perfect predictions of the future on Life. The other manual tells us about persisting entities - it generates very good predictions provided there is no noise. Ascending to the level of persisting entities carries some risk, it is not perfect after all, but also carries major benefits especially in the realm of computing time and complexity. Dennett suggests that the patterns are there for the picking up and will be if the right algorithm is hit upon. Basically he runs the Quinian line that when you have radical translation there is no fact of the matter between two translation manuals as to which is the right one - the so-called indeterminacy of translation.[34] As Quine would have it not only is there no fact of the matter but asking the question itself makes no sense because there is no conceivable way of telling which is the right translation. Dennett's position on intentional states is that they are just elements of one among many translation manuals of human behaviour. It makes no sense to ask whether beliefs (or gliders) really exist, or analogously whether the *intentional stance* is the correct explanation of reality. Our ontological commitment to these entities depends on the way they function as terms in our language. Beliefs cannot play as wide a role as material objects can and so we are less committed to them than we are to the material world. But just because we are less committed to beliefs than we are to material things it does not mean that beliefs do not exist or are not objectively there.

---

[34]Quine uses the following example. An anthropologist is translating a native's language. The native utters the sentence "Gavagai" as a rabbit hops past. Now the anthropologist translates the sentence as "rabbit", or "lo a rabbit". But Quine points out that 'gavagai' may refer to rabbit body parts, or time segments of rabbit, or fluffy white animal equally plausibly. He provides a long discussion of this in *Word and Object*, but the gist needed to understand my point here can be found at pp. 51-4.

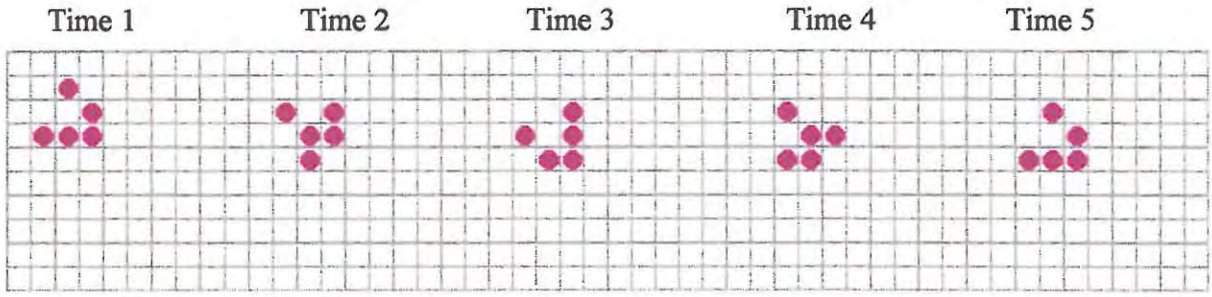

Figure 1: A Glider takes five ticks to advance itself by one space

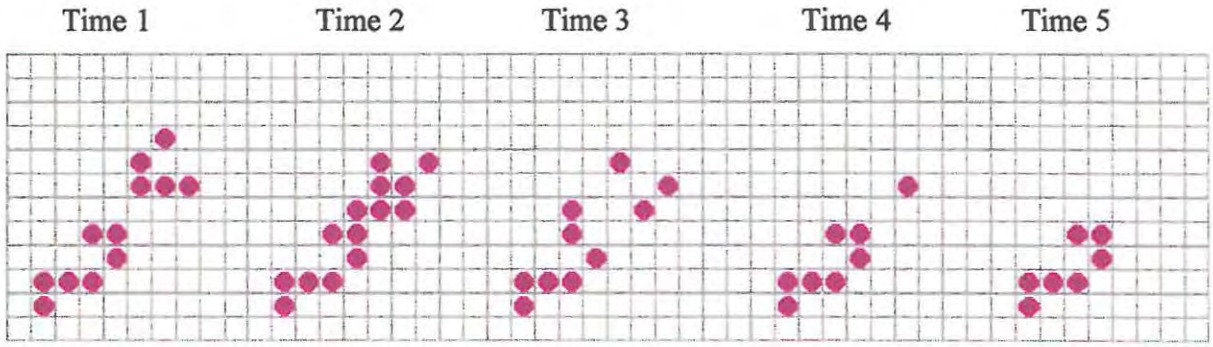

Figure 2: An Eater consumes a Glider as the Glider touches it

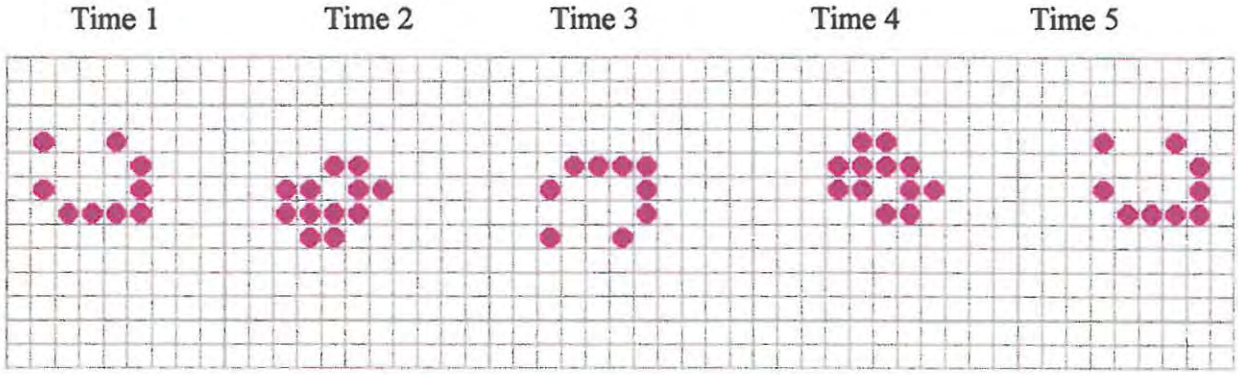

Figure 3: A Ship slowly makes its way across the grid

Note that we can choose to adopt a perspective where there are Gliders, Eaters, and Ships in Life. We can also adopt a perspective where, by applying a simple algorithm, cells are merely either on or off depending on their previous state and of that of their surrounding cells. Adopting such a perspective allows us to predict exactly what the next state will be. (Graphics generated on PageMaker from WinLife by John Harper)

What is abundantly clear from this elucidation of Dennett's position is that for him intentional talk is not really eliminable. When we are faced with a mechanistic system we can explain its behaviour in at least two ways, depending on what translation manual we are using. So the fact that there is a mechanistic explanation does not mean that beliefs and desires cease to exist. Of course, it is clear that in such a case we do not *have to* use intentional jargon to explain the behaviour of the system, so it is eliminable in a sense. But because of the Quinian notion that there is no fact of the matter as to which way something is to be translated, the intentional translation cannot be rejected altogether. It is as legitimate as the mechanistic one. Basically Dennett's position is one where whether we use a mechanistic explanation or an intentional one is a matter of at which joints we choose to carve nature. The idea is something like the notion that mechanistic explanations carve digit by digit, while intentional explanations take a limb at a time. Now, since intentional states on Dennett's account do obviously occupy some sort of objective position and are not totally eliminable the question remains as to whether there is a real difference between Dennett's conception of the intentional states and the folk conception.

The grounds of an answer to our question have already been laid. We have seen that the realism of the folk conception of intentional states is grounded in the fact that on that conception we have to advert to such states in order to explain the actions of an intentional system. In other words, from the folk perspective we are never faced with a case of indeterminacy of translation. If there is a mechanistic explanation of a system's behaviour then we must use the mechanistic translation manual. If, on the other hand, there has to be an intentional explanation of the system's behaviour then that is the translation manual we must use. So on the folk conception there is a fact of the matter as to how to interpret behaviour. This perspective can be contrasted with what we have seen regarding Dennett's position. Namely that we do not have to employ intentional terms to explain behaviour on Dennett's account. Furthermore it is apparent that although intentional states can be said to exist on Dennett's conception, they really depend for their existence on our adopting a particular interpretive stance toward behavioural systems. Pushing the contrast to its limit, we see that in a plausible sense the *intentional systems* as picked out be the *folk stance* can be said to really believe and desire while those of Dennett's *intentional stance* would be better

characterised as merely behaving *as if* they believe and desire. Given this outcome I propose that we think of the *folk stance* as being an intentional stance proper, and that Dennett's *intentional stance* is better understood as a *pseudo*-intentional stance.

6.3 Explanatory Compatibilism Again

The notion of explanatory compatibilism, we have seen, is that mechanistic and intentional explanations of a system's behaviour can coexist. Now that we have a deeply enriched understanding of Dennett's attempt to secure such a position it is clear that his sort of intentional explanation is *indeed* compatible with mechanistic explanation. This outcome is not, after all, surprising in the least, since Dennett standardly employs examples of known mechanisms to illustrate his *intentional stance*. What we have also seen, however, is that Dennett's sort of intentional explanation is not the only sort of intentional explanation. The alternative conception, called the *folk stance*, is one which truly has action as opposed to mere behaviour at its heart. In essence it focuses on a much narrower sphere of behaviour than Dennett's stance does. It only pertains to behaviour that is brought about because of a certain set of intentional states and which the system, or agent, wills. The question that remains is whether Dennett's move could possibly extended to this alternative.

I want to suggest that the preceding discussions of *method-dependent rationality* and the folk conception of intentional states have shown beyond all doubt that Dennett's compatibilism does not, and cannot, extend to include intentional explanation from the *folk stance*. We have seen that the folk idea of rationality is such that it explicitly excludes systems whose behaviour can be mechanistically explained from being rational. This in itself is enough to show that Dennett's compatibilism could not include folk intentional explanations. His compatibilism is, after all, at least partly generated from a conception of rationality that is specifically tailored to include mere mechanisms. That conception of rationality, *method-independent rationality*, is purely a matter of external assessment of whether a system's behaviour fits with its putative beliefs and desires. But the case against extending Dennett's compatibilism can be made even stronger.

It has been established that a further element to Dennett's compatibilism lies in his understanding of what intentional states actually are. In other words, his compatibilist position is also legitimated in part by his understanding of the existential status of the things designated

by the terms which populate intentional explanations in general. His understanding, we have seen, is one where notions such as belief and desire depend for their existence mainly on their being elements to a pattern which we discern when we adopt his *intentional stance*. The fact is, however, that this conception ultimately makes intentional states merely a matter of interpretation and consequently makes Dennett's position one of straight instrumentalism. I have called these *pseudo-intentional states* because of the lack of a robust realism associated with them. Now it is obvious that this conception can be extended to include real (as opposed to *pseudo*) intentional states and so it may seem that Dennett's conception is not in any difficulty. But the folk understanding of intentional states with its attendant realist perspective is such that these intentional states *must* feature in the causal story behind the system's behaviour. In fact, this element of the folk conception is a major factor in the realism of the position. Now Dennett's *pseudo*-intentional states clearly do not occupy the same status. Pseudo-intentional states do not *have to* feature in the causal story of a system's behaviour.[35] This consideration would suggest that the two conceptions of intentional states are actually vastly different and that Dennett's *pseudo-intentional stance* must remain forever apart from the *folk stance*. As a consequence of this unbridgeable divide between the two positions it follows that even though Dennett has shown his kind of intentional explanation to be compatible with mechanistic explanation he has not made the case for the folk intentional explanation. With this conclusion firmly in mind we can now return to the issue of moral responsibility.

6.4 Two Intentional Stances and the issue of Moral Responsibility

Earlier on I discussed the conditions surrounding our typical ascriptions of moral praise and blame. What I wish to investigate now is which of the two 'intentional stances' is best suited to our judgements of moral responsibility. My claim will be that Dennett's position is in fact ill-suited to judgements of moral responsibility and that the folk perspective fares much better. Putting this together with the case I have already established for saying that Dennett's compatibilism does not extend to the *folk stance* I suggest that Dennett is unable to secure the

---

[35]There is one exception to this of course. When the system is not a mechanistic one the *pseudo*-intentional states are ultimately the system's real intentional states. In such a case, obviously, a mechanistic explanation would not be available and the intentional explanation would be the only explanation.

compatibilism of mechanism and responsibility. The argument is simple and decisive. If our judgements of moral responsibility are made from the *folk stance* and they are not made from Dennett's '*pseudo*-intentional stance', then Dennett is unable to secure the case for the compatibilism of mechanism and responsibility. Our judgements of moral responsibility *are* made from the *folk stance* and not from Dennett's position. Therefore Dennett cannot, and has not, secured the case for the compatibilism of mechanism and responsibility. Basically I plan to begin with a brief recapitulation of what is involved in our judgements of moral responsibility. I then show that the *folk stance* captures the sense of what is involved very neatly. Finally I show that Dennett's position does not achieve this.

The notion of responsibility only really becomes an issue when we believe that we are faced with a system that is an agent. As long as we do not believe a system to be an agent we do not ascribe responsibility in any but the most tenuous sense. So, for example, we *might* say that the wind is responsible for the roof damage, but we would only mean that the thing which caused the roof to become damaged was the wind. We would not mean that the wind *did* the roof damaging in the sense of action. Recall, for a further example, our discussion of Josephine's breaking the Ming vase because of an epileptic seizure. She was not judged responsible because she did not act. What, then, is involved in the notion of agency? A major consideration is that for a system to be considered an agent it must be such that we have to give and intentional explanation of its behaviour if we are to explain its behaviour at all. That is, if a system is to be regarded as an agent we must believe that its intentional states actually play a role in the causal factors of its behaviour.

A second element to our ascription of responsibility lies in the idea that we tend not to ascribe responsibility unless we believe that the system believes that it had some choice in the matter of what it was going to do. If, for example, we believe that the system did what it did because it believed that it could not have done otherwise we would generally not assign responsibility. So when a system believes that it is coerced, or is perceived to believe that it has been unable to refrain from doing what it did, or when there is seen to be any *force majeure* propelling behaviour forward, we do not hold the system responsible. This point is highlighted if we consider an example. It may seem that responsibility is judged on *what is* and is not simply a matter of what the subject believes, but this is wrong. Consider that when I ask you why you did not turn off the stove and you reply that it was because you did not

believe that it was on, I cannot hold you responsible for your lack of action. I might hold you responsible for your belief if, say, I told you that it was on, but that is now a different issue and does not pertain to your behaviour. So we see that when this element is not satisfied we do not ascribe responsibility. Except, of course, in the tenuous sense expressed above where the wind was responsible for damage. This sense is most certainly not that of moral responsibility. Linked to this element of our responsibility judgements is the notion that we can only judge responsibility when the cause of a system's behaviour lies in the reasons it has for behaving and not some chain of antecedent physical events. The reason we have this notion is that when we judge responsibility we are ultimately claiming to have identified the culprit responsible for a certain set of circumstances. When behaviour is caused by a set of antecedent physical events we cannot legitimately lay claim to have identified the perpetrator of the behaviour if we halt our search only at the level of the system itself. In fact the ascription of responsibility in such a case is, I have suggested, merely a case of settling on a scapegoat and is an arbitrary affair.

The question before us now is whether the *folk stance* is suited to the notion of ascribing responsibility or not. I believe that the case for claiming that it is has already been made perfectly clear. We have seen that the *folk stance* is such that it aims specifically at behavioural systems which are considered to be agents. These systems are such that we have to provide an intentional explanation of their behaviour if we are to explain it at all. In adopting the *folk stance*, furthermore, we ascribe a particular conception of rationality wherein the system's belief that what it does is up to it plays a major role. Finally we have seen that the *folk stance* involves a conception of the origin of behaviour which places the cause of the behaviour with the system itself rather than with some chain of antecedent physical events. This shows clearly that the *folk stance* is particularly well suited to the ascription of responsibility.

Now, is Dennett's *intentional stance* up to the task? Can it be said to be suited to the ascription of moral praise and blame? Once again I believe that the answer has already been clearly detailed in our preceding discussion. Dennett's position has been shown to be such that we most certainly do not have to employ intentional explanations of intentional systems in general. It is also such that choice does not feature at all. A system does not have to believe that what it does is up to it in order to qualify as an intentional system for Dennett. Finally

Dennett's stance explicitly caters for systems whose behaviour is mechanistically caused and which thus does not originate with the system itself. These show that Dennett's position is obviously not well suited to judgements of responsibility. This conclusion is further bolstered by the fact that firstly we have seen that the *folk stance* is well suited to judgements of this kind and secondly by the fact that we have seen that there is an unbridgeable gap between the two stances.

It should be perfectly clear now that Dennett's *intentional stance* is not the position from which we judge responsibility. Furthermore it is particularly ill-suited to being thought of as such a position. Now the fact that Dennett has made a case for the compatibilism of his kind of intentional explanation and mechanistic explanation can be seen to have no impact on the question of mechanism and responsibility. Dennett's case does not extend to include the *folk stance* and it is from the *folk stance* that we actually make our judgements of moral responsibility. The upshot is then, that Dennett fails to reconcile mechanism and responsibility. The main reason for this lies in the fact that his attempt literally misses the mark in that it focuses on the wrong sort of intentional explanation.

7.0 Concluding remarks

The time has come to take stock. I have covered a lot of ground and now wish to consolidate the process by providing a brief overview of the argument as it was developed. In essence I am now going to give a step by step run through the investigation highlighting salient issues and tying off any loose ends.

The first step was to basically draw out the intuitions regarding moral responsibility. At that stage we saw that there is a strong intuition that the ascription of responsibility is only legitimate when we have to give an intentional explanation of the behaviour. I went on to show that there is a further intuition that intentional explanations are undermined by mechanistic explanations. We saw that there appears to be a strong *prima facie* case to this effect. One way that the resultant explanatory incompatibilism could be avoided, I suggested, would be to adopt a form of reductive materialism. This move would identify intentional states with brain states and thus effectively rule out the issue of incompatibilism. This, of course, is not the route that Dennett takes and so we turned to examine that. From Dennett's perspective we saw that there seemed to be an equally plausible case *for* compatibilism. He systematically depicts mechanistic systems from an intentional perspective and no violence seems to be done to the intentional characterisation by the fact that the systems are quite obviously *mere* mechanisms. What this showed however, was only that mechanistic behaviour can be intentionally characterised and not that intentional behaviour can be mechanistically explained. This is obviously the other half of the battle in reconciling the two vying explanatory strategies.

With this in mind we found that Dennett develops a conception of intentional explanations that includes a specific understanding of what rationality is, and what ontological status is to be assigned to the intentional states. The rationale behind this move on Dennett's part is obvious. What he seeks to show is that behaviour that counts as intentional in this account is not committed to the truth or otherwise of mechanism. Basically the move seeks to show that there is *no* reason why intentional behaviour cannot be mechanistically explained. The place to look for shortcomings in Dennett's account clearly lay in this very broad *method-independent rationality* he uses and, of course, his position on intentional states. I argued that *method-independent rationality* just does not do justice to our requirements regarding action

and agency. I showed that there is a different sort of rationality, *method-dependent rationality*, involved in our understanding of agency and that it is vastly different to Dennett's. This, I claimed, is indicative of the idea that there are in fact two different sorts of intentional stance and thus two different types of intentional explanation. It also means that Dennett only secures a compatibilism for one of the types.

To bolster the case against Dennett I turned to the issue of intentional states. That is, I set about showing that aside from the two different understandings of rationality employed in the two stances, there are other irreconcilable differences between the two. I showed that Dennett's *intentional stance* really defends a conception of intentional states that is interpretationist or instrumentalist while the alternative, which I called the *folk stance*, demands a strong realism. I went on to indicate that it is actually from the *folk stance* that we make our judgments of responsibility. Beside the fact that the *folk stance* excludes the possibility of mechanistic systems being truly characterised as intentional systems by its very nature, I suggested that seeing as Dennett clearly reconciles the non-relevant type of intentional explanation, namely his own, with mechanism he ultimately fails to reconcile mechanism and responsibility.

In the final analysis then we see that Dennett's *intentional stance* is better characterised as a *pseudo-intentional stance*. It exploits a conception of rationality that is so broad that it enables all sorts of systems to qualify as intentional systems, many of which we would not count as intentional systems from the folk perspective. The conception of intentional states that goes with Dennett's stance has also been found wanting. It is clear that Dennett does secure explanatory compatibilism between his type of intentional explanation and mechanistic explanation but he has not gone far enough. In particular he has been unable to reconcile genuine intentional explanations with mechanistic ones. In fact it seems unlikely that this is even possible. As such we see that Dennett has failed to allay our fears for the notions of moral responsibility and intentional explanation in the face of popular belief in the idea that we will eventually uncover the mechanistic nature of our behaviour. I am not persuaded by this idea and hope to see a coherent and plausible conception of agent-causation developed that secures the matter of our agency and the possibility of our being held morally responsible for what we do. This, however, is a major undertaking and must, for now, remain a possibility for future consideration.

Bibliography

Books and Articles referred to in text:

1 Chisholm, R (1966) "Freedom and Action" in Lehrer, K (ed) *Freedom and Determinism*, (New York-Random House)

2 Clarke, R (1993) "Toward A Credible Agent-Causal Account of Free Will" *Nous*, 27

3 Dahlbom, B ed. (1993) *Dennett and his Critics*, (Oxford-Blackwell)

4 Davidson, D (1980) *Essays on Actions and Events*, (Oxford-OUP)

5 Dawkins, R (1989) *The Selfish Gene*, (Oxford-OUP)

6 Dennett, D (1978) *Brainstorms*, (Hassocks, Sussex-Harvester Press)

7 Dennett, D (1984) *Elbow Room: The Varieties of Free Will Worth Wanting*, (Oxford-Clarendon Press)

8 Dennett, D (1987) *The Intentional Stance*, (Cambridge, Mass-MIT Press)

9 Dennett, D (1991a) *Consciousness Explained*, (Boston, Mass-Little Brown Books)

10 Dennett, D (1991b) "Real Patterns" in *The Journal of Philosophy*, 88

11 Hume, D (1977) *An Enquiry Concerning Human Understanding*, (Indianapolis-Hackett Publishing Company)

12 Kim, J (1993) *Supervenience and Mind: Selected philosophical essays*, (Cambridge-Cambridge University Press)

13 MacIntyre, A (1957) "Determinism" in *Mind*, 66

14 Malcolm, N (1968) "The Conceivability of Mechanism" in *The Philosophical Review*, 77

15 Nagel, T (1986) *The View From Nowhere*, (New York-OUP)

16 Nozick, R (1981) *Philosophical Explanations*, (Oxford-Clarendon Press)

17 Quine, W.V. (1960) *Word and Object*, (Cambridge, Mass-MIT Press)

18 Stich, S (1981) "Dennett on Intentional Systems" in *Philosophical Topics*, 12

19 Taylor, R (1963) *Metaphysics*, (Englewood Cliffs, New York-Prentice Hall)

20 Watson, G (1982) *Freewill*, (Oxford-OUP)

Other works consulted:

1 Brooks, D (1995) "Corporate Minds" - unpublished staff seminar University of Cape Town

2 Clark, A (1994) "Beliefs and desires Incorporated" in *The Journal of Philosophy*, 91

3 Gert, B (1990) "Rationality, Human Nature, and Lists" - *Ethics*, 100

4 Long, D (1979) "Agents, Mechanisms, and Other Minds" in *Body, Mind, and Method*, D. Gustafson & B. Tapscott (eds.)

5 McGinn, C (1989*) Mental Content*, (Oxford-Blackwell)

6 Nozick, R (1993) *The Nature of Rationality*, (Princeton, New York-Princeton University Press)

7 Pollock, J (1989) *How to build a person: a prolegomenon*, (Cambridge, Mass-MIT Press)

8 Sheridan, G (1983) "Can there be moral subjects in a physicalistic universe?" in *Philosophy & Phenomenological Research*, 43

9 Waller, B (1993) "Responsibility and the Self-made Self" - *Analysis*, 53

10 Wilkes, K.V (1994) *Real People: Personal identity without thought experiments*, (Oxford-Clarendon Press)