Why mental explanations are physical explanations[1]

Julian Mark Jackson

Department of Philosophy, University of the Witwatersrand, Private Bag 3, WITS 2050, Republic of South Africa

Received August 1994; revised October 1994

In this article the author argues that mental explanations of behaviour just are physical explanations of a special kind. The ontological reduction of mental events to physical events is endorsed. It is argued that explanations of behaviour citing mental events are not mysterious because they designate events with normal, physical causal powers. Mentalistic terms differ from physicalistic ones in the way in which they specify events, the former doing so by citing extrinsic properties, the latter intrinsic properties. The nature of explanation in general is discussed, and a naturalistic view of intentionality is proposed. The author then shows why *epistemological* considerations rule out the elimination of 'mentalistic talk' from explanation of behaviour.

In hierdie artikel betoog die skrywer die siening dat geestelike verduidelikings van bedrag bloot spesiale soort fisiese verduidelikings is. Die ontologiese redusering van geestelike gebeurtenisse tot fisiese gebeurtenisse word onderskryf. Daar word betoog dat geestelike verduidelikings van gedrag nie misterieus is nie, aangesien hulle wys op gebeurtenisse met normale fisiese kousale kragte. Mentalistiese terme verskil van fisiese terme slegs ten opsigte van die manier waarop hulle gebeurtenisse spesifiseer, eersgenoemde op grond van ekstrinsieke eienskappe, laasgenoemde op grond van intrin-sieke eienskappe. Die aard van verduideliking in die algemeen word bespreek, en 'n naturalistiese beskouing van inten-sionaliteit word voorgestel. Die skrywer wys dan waarom *epistemologiese* oorwegings die uitskakeling van 'mentalistiese gesprek' buite rekening laat in verduidelikings van gedrag.

1. Introduction

In this essay I will examine certain difficulties that arise for the explanation of intentional action. According to a common-sense view we can explain an agent's actions by giving his or her reasons. Having reasons involves being in certain mental states, and these mental states cause action. However, mental states have several properties which distinguish them from physical states. First, they appear to have representational properties. Thoughts and other mental states are *about* things in the world, they *represent* the world in virtue of having a certain content.[2] And second, mental states have an immediate experiential character. One might say that they have a subjective 'feel', which is essential to their being mental states. In this essay will only be concerned with the former aspect of mental states, and it raises an interesting difficulty. If we assume an externalist's position on how mental states acquire content we say that content is an extrinsic property of mental states. But further intuitions seem to commit us to the view that only the intrinsic properties of states are causally relevant. Does this show that mental properties are not causally efficacious?

In the course of answering this question I will explore the relationship between physical explanations and what I term 'mentalistic' explanations. In particular, I will examine whether physical explanations can replace mentalistic ones. In the end I will forward a position which will probably put me in the reductionist camp. I will say that mental explanations just are physical explanations of a special kind. And I will say that only epistemological and pragmatic considerations stand in the way of reduction. If we were omniscient, if we knew what the reduction base for mental states was, there would be no obstacle; ontologically the road is clear. But I will insist that in view of how little we know about the world we cannot reduce mentalistic language. And so, not much will depend on whether I am advocating a reductionist position or not.

2. Reasons as causes

In broad terms, intentional action is simple to define. 'Intentionally' means willingly, or purposefully. Therefore, we are concerned with 'willed' or 'wilful' action. Perhaps the concept 'action' already contains that meaning, in some cases at least: 'Jeff acted' seems equivalent to 'Jeff acted willingly'. Though, in ordinary language, saying 'Jeff acted, but against his will' need not be construed as a contradiction. But we do not need to complicate matters. For the purposes of this essay, we will assume that action is always intentional.

Defining 'action' accurately has proved to be a notoriously difficult task. Again, we need not get involved in details to be able to see the difficulties concerning mental to physical causation. But some background may help to illuminate the issue at hand. We must contrast acting with being acted upon, as happens, for example, when the current in an alpine stream drags somebody away despite his best efforts to stay on his feet. Behaviour resulting from outside forces exclusively is not action. But internal causes are not sufficient for action, either. Consider the ongoing processes of blood-circulation and digestion. These occur thanks to internal causes, and yet we prefer to say that something is happening to us, rather than that we are acting. The obviously missing feature for action is that it needs to be caused by internal psychological states, like beliefs and desires. But even this is not enough. Action must be caused by mental states in the right way. Illustrating this, there is the well-known example of the mountain climber who wishes to kill his unsuspecting companion who is in the unfortunate position ot dangling from a rope, which is being held by our would-be killer. Contemplating the thought of letting go disquiets our man so much that he momentarily loses control of the rope. Thereby he accidentally accomplishes his premeditated aim. His letting go of the rope was caused by his thoughts. Yet we should resist saying that he acted, because the outcome was accidental. Non-accidentality is extremely difficult to capture in this context. Part of its significance, however, is that for action, the reason why the man let go of the rope ought to be *his* reason. Action must be caused 'non-deviantly',[3] as planned.

We need not delve into the concept of action any further to see that it is closely tied to an agent's choice, and that acting for reasons is as common as action itself. Aristotle's meaning is thus quite clear when he says (Heil, 1992: quoting Nicomachean Ethics,1139a: 31–33).

> 'The origin of action — its efficient, not its final cause — is choice, and that of choice is desire and reasoning with a view to an end.'

Action arising in just this manner is commonly observed. Examine the following case: The barman, Lofty, at the local bar stands behind the counter, looking sceptical. He is scrutinizing a swaying customer, who, sitting opposite, can hardly stay on the stool. Lofty is listening to the drunk, who is muttering away: '"This is the limit. They have started to water down the drink. I'm injured and hurt. To think that my patronage means nothing to these people. Well, that's it. I'm leaving. If only I could find the car keys...' Lofty knows the scene; it happens every Friday night. The difference is that tonight the keys Thirsty is looking for belong to Lofty's car. And Thirsty is borrowing Lofty's car as a result of Thirsty's navigatory misfortunes last Friday. Lofty faces what is known in metaethical circles as 'The Barman's Dilemma': To sacrifice a car to a drunk, or save it from an accident. Lofty considers carefully, then makes his choice, based on this reasoning: 'If I let this drunk take the keys he will crash the car.' As a result, he quickly snatches the set of keys which Thirsty has just retrieved from the bottom of his beer-mug.

There is no difficulty in seeing that Lofty acted, and that he acted for (good) reasons. The reasons for the action are *Lofty's* reasons because he willingly, intentionally, *brought about* the outcome, and in the right sort of way. As Aristotle would have put it, *Lofty's choice caused the outcome*. Thus reasons can be causes of action. I put this forward as a common-sense view, which is no more remarkable than the notion that mental events, thoughts, cause behaviour (Kenny 1970: 142):

> 'Everyone feels that he is a single person with both body and thought so related by nature that the thought can move the body and feel the things which happen to it.'[4]

Accordingly, it seems plausible to suppose that what thoughts are about matters to how thoughts 'move' the individual. The content of somebody's thoughts plays a role in explaining behaviour. Lofty's beliefs about Thirsty are causally relevant just because of the content they have. Lofty thinks, 'Thirsty is drunk'. This is different from 'Jeff is drunk', which, having a different meaning, generates different action. If a thought referring to 'Thirsty' was about the local parish priest who has signed the pledge and sticks to it, then Lofty's 'reasons' for taking the keys from Thirsty would have been quite without explanatory power. Conversely, it is just because 'Thirsty' means what it means that Lofty's reasons *are* reasons,

So far, there seems to be no reason to think that there is any difficulty in conceiving of reasons as causes for action. However, there are certain considerations which make the idea appear puzzling.

3. Intrinsic properties and externalism

The causal relevance of what thoughts are about, or what their content is, points to an interesting dissimilarity between reasons as causes and causes of the sort we are familiar with from the natural sciences. This is best illustrated by an example. When Gomez Addams whacks a golfball off his balcony with a Driver and breaks the judge's bedroom window a few hundred metres away, we must suppose that the ball's having a certain weight, moving at a certain speed and possessing certain shape are responsible for its ability to break windows. Entirely irrelevant seems to be the fact that the ball was manufactured in the United States by a company called 'Titleist', or its property of having been hit by Gomez and not his brother Fester. These seem to be properties of an entirely different 'type'. The former appear to be straightforward 'physical' properties, whereas the latter are relational properties. Physical properties are 'here-and-now' intrinsic properties, and it seems that golfballs can do damage to windows solely in virtue of intrinsic properties of themselves and their trajectories. Relational properties are extrinsic. The property of 'having been hit by Fester' depends on the interaction of the ball and Fester, who is an independent entity.

I will soon argue that there are good reasons to believe that mental states' possession of content is not one of their intrinsic properties. This will give rise to a puzzle. If content plays a role in explaining behaviour, and if content is not an intrinsic property, it will (apparently) have to do causal work in a rather different way than the intrinsic properties of a golfball and its trajectory. But before developing this argument, let us return to the common-sense claim, that only intrinsic properties of objects or events are responsible for their causal powers, and that relational, extrinsic properties do not affect the causal process.

We have very strong pre-reflective convictions about the idea that only intrinsic properties are causally significant. In situations with similar causal powers we assume a similarity of 'internal structures' which explain similar outcomes. Positing causes just is attributing a similarity of internal structures. Imagine somebody examining all the squeaking hinges in his house. Every squeaking hinge is dry. He concludes that the cause of squeaking in hinges is dryness. This inference is made even though in principle it is possible that some hinges always squeak, even when wet. The inference is driven by the conviction that similarities of outcomes are due to internal similarities. And it works the other way round, too. If situations are intrinsically indistinguishable we assume that their causal powers are also indistinguishable. Two intrinsically indistinguishable golfballs are taken to have indistinguishable causal powers.

However, there are cases when we might be inclined to disagree[5] with the common-sense view. Heil offers the following example: Two balls, A and B, of equal volume but unequal mass are dropped from the same height, at the same rate of acceleration, into a sandpit. They make different imprints, one inch deep and two inches deep respectively. The phenomenon is due to the unequal masses of the balls. These appear to be intrinsic properties. But the heavier ball (B) seems also to have the added property of being capable of making an imprint one inch deeper than ball A. This property is surely dependent not only on its intrinsic properties but also on ball A's intrinsic properties. In other words, an extrinsic property of ball B is being heavier than ball A. Where does this leave our notion that relational properties are not causally relevant?

Heil argues[6] that we should distinguish between an object's 'broad' and 'narrow' causal capacities. Narrow causal capacities are shared by intrinsically identical objects, broad causal capacities are not necessarily shared. On Heil's view, broad capacities are environment-dependent. Imagine me consecutively dropping two intrinsically identical balls into a sandpit. While I drop the first one I sing, 'Swing low...' While I drop the second, I sing, 'Get high ...' In Heil's view, the first ball can be described as having two causal capacities, of interest to us: it has the capacity to produce an imprint one inch deep, and it has the capacity to produce an imprint after having been dropped by me while singing 'swing low ...' The second ball will have the same narrow causal capacity, for it will also produce a one-inch imprint. But it lacks the broad capacity to be dropped while I sing 'swing low ...'.

Instead, it has the capacity to produce an imprint having been dropped by me while singing 'Get high ...'.

But I think there is a problem with this distinction between narrow and broad capacities. If I had sung 'swing low ...' while I dropped the second ball, undoubtedly the ball would still have dropped to the ground. What then is the sense in which we say that it lacked the capacity to be dropped while I sing 'swing low ...'? What one ought to say is that it lacked that capacity in this world, given the fact that only ball A is dropped while I sing 'swing low ...'. One should say, 'in this world, it could not exercise a capacity to be dropped while I sing 'swing low ...'. In this sense it does not have that capacity. And therefore, broad capacities are context dependent. So long as the external circumstances for balls A and B are different, they will have differing broad causal capacities. With this addition the Heilean distinction can usefully be maintained.

Whether we 'pick out' the broad or narrow causal capacities of objects depends on our interest. For example: Thirsty's wife may develop an irrational dislike for Thirsty when he gets home on Friday nights. In order to express her dislike in a way that is unlikely to be misunderstood, she is in the habit of taking a rolling-pin to his backside. Why does she choose the rolling pin? Because it possesses a vital property — that of permitting her to communicate with Thirsty clearly. This is a complex, relational property. Though it is the property of the rolling-pin that she most values it is not an 'essential' property. Notice that the rolling-pin gets this relational property thanks to its intrinsic properties of having a certain mass, shape, etc. Indeed, anything with relevantly similar intrinsic properties as the rolling-pin would do the trick. This indicates that relevant, narrow, causal capacities are often contingently[7] identified via broad capacities. But this should not distort the fact that we are often after the relevant narrow capacity. In a footnote, Heil makes the general point clear:

> 'How we choose to identify features of the world depends, not only on the features we want to identify, but also on the context in which the identification is made and on innumerable pragmatic factors. I may identify a particular electromagnetic wavelength, for instance, by means of the description, "my least favourite colour". Here, as elsewhere, it is important to distinguish a *mode of description* from *whatever is picked out* by the description' (last sentence my ital.).[8]

It seems clear that what matters independently of how we pick out a relevant property is the property itself — what is picked out by the description. And that will be, where causation is concerned, an intrinsic property.

We have the first part — only intrinsic properties are causally relevant. The second part of the story is the solution to this problem: is content an intrinsic or a relational property? The philosophical world is divided into two camps: internalists and externalists.[9] The former, who are arguably losing the battle,[10] think that content is grounded in one thing only — the internal properties of the agent's mind. The latter believe that content also depends on conditions outside the agent. I want to say a little about both views without going into great theoretical detail. First internalism. The basic tenet of internalism is that the content of mental states depends solely upon intrinsic features of the mind.[11] Thus content is said to 'supervene' entirely on internal states, or be realized by them. A consequence of this view is that,

> '... if we hold these intrinsic features constant and vary the context in which they occur, we may alter the truth-value (or satisfaction-value) of particular attitudes, but we do not thereby alter their satisfaction-conditions or their content. I believe truly that there is a sheet of paper in front of me. Suppose an evil demon causes the paper to vanish while simultaneously inducing me to hallucinate a sheet of paper. I may now believe falsely that there is a sheet of paper in front of me, but the content of my belief is unaltered: It is still a belief about a piece of paper, a belief that is true if and only if there is a sheet of paper in front of me' (Heil 1992: 23).[12]

The essential aspect of internalism is that the content of mental states supervenes exclusively on intrinsic[13] features of the agent, which means that content can supervene on internal states independently of the existence of other objects. Accordingly, internal states could, in principle, realize contentful states about objects which have never existed. For example, suppose 'grog' means an object which materializes out of the matter of my desk when the sun stands in line with the earth and the centre of the universe. Suppose also that this has never happened, but that it will, in exactly two days. According to internalism a person could have thoughts about grog even if he or she has never encountered the stuff, before it materialized. Perhaps internalists would insist that the agent would have to be aware of the thought, like, 'Hey, I'm thinking about stuff which I've never seen before'. When grog finally appears, and the person thinks about it, we will find that whatever internal states were there when the person thought about grog before grog appeared, will exactly resemble those that occur when the person makes the acquaintance with grog and thinks about it. Again, perhaps internalists would insist that the agent has to be aware of the similarity if it is to be a genuine similarity, like, 'this is the stuff I was thinking about the other day'.

Externalists believe that a mental state has the content it has not only in virtue of its intrinsic properties, but also in virtue of its causal/historical properties. Thus, content is said to depend also on context, including the circumstances of agents.

> 'In altering agents' circumstances, we may vary the content of their intentional attitudes. The character of those attitudes depends not solely on the intrinsic features of agents, but also on relations those agents bear to extrinsic states of affairs (Heil 1992: 24).[14]

Externalism takes seriously a difficulty with internalism: what is the mysterious means by which internal states turn out to be *about* something? If we resist positing a kind of pre-established harmony, or a divine order, it is difficult to see why a state's being this way or that should, by itself, make any difference to what it means. Externalists hope to solve the problem by arguing that what something means depends on the context within which it occurs. For conventional signs this is an undisputable claim. Consider the meanings of gestures. What does raising my hand mean? Intrinsically it has no meaning at all. Only when we add the context can we determine a meaning, like that I raise my hand in class in a society in which raising one's hand in class is usually taken to mean, 'I want to speak'. In a different context, as when I wake up in bed in the moming, raising my hand may mean nothing at all, I may be stretching. And at Sotheby's raising my hand may mean 'I bid'.

According to externalism, mental content must be explained similarly. Take my thought about 'Earl Grey Tea'. In a typical situation it might be caused by Earl Grey Tea. Admittedly, were my thought about 'Five Roses Tea', my internal states realizing that thought would be somewhat different. But when we consider the matter it seems as if what matters to 'Earl Grey Tea' meaning Earl Grey Tea and not Five Roses Tea is not what the internal state is, so much as how the state was caused. Since 'Earl Grey Tea' was caused by Earl Grey it means Earl Grey.

One side-remark: In explaining the externalist's view that the content of a state depends on its causal connections[15] I do not mean to suggest that each individual token-state gets its content that way. There is good reason to think that what matters for the content of a state is that the state belongs to a type which is normally caused in a certain way's. Then each token-state has the content it has in virtue of its belonging to a certain type-category, and the particular causal connections possessed by the token-state do not determine its content.

The issue is relevant to representation in general. I do not want to get into a detailed discussion of the requirements for representation. Let it suffice to say that it is difficult to imagine how the causal connections of token states could, alone, give them content. For if it were so, it would not be possible to say why a certain experience, say the viewing of a sunset, has a 'viewing a sunset' content. Viewing a sunset just is a case of detecting certain physical properties of the sun via our neural network. The reflection of light, the working of our optic nerve, neural firings, these are all part of the causal chain causing the experience. And so, why does our experience not have a 'configuration-of-atoms-content, a pattern-of-light-content, or a neural-firings content'? (Pendlebury, p. 7) Why is the sunset relevant, and not the rest? The answer is, loosely put, that not all parts of the causal history typically cause an experience of that *type*. Though neural firings always are part of sensation, and though *this* neural firing does partly cause my sunset experience, *this* neural firing does not *normally* cause a sunset experience. And so we need to differentiate those aspects of the causal history which are generally applicable to similar cases of experience. So why is the sunset the most significant part of the causal history, for the purposes of determining the content of the experience?

> 'First, the experience belongs to a similarity class of visual experiences the members of which are typically caused by the presence of [a sunset], but which are not typically caused by any single configuration of atoms, light rays, or neural firings. Any given experience in this similarity class [is] of course caused by some configuration of atoms, light rays, or neural firings, but there is no single configuration of any of these types which is normally involved in the causal origins of experience in the relevant class in the way that the presence of [the sunset] is. Second, the visual experiences in this class typically cause or are apt to cause behaviour which is ecologically appropriate to the presence of [a sunset], but not especially to the presence of a given configuration of atoms, light rays, or neural firings' (Pendlebury, p. 7, substitutions mine).

If a particular experience is to represent a sunset, it is not enough if it was caused by a sunset. For that same experience was also caused by certain neural firings, and it does not represent neural firings. This shows that the causal connections of a particular experience do not determine what it represents. Rather, what matters is that a class of relevantly similar experiences (an experience-*type*) is *normally* causally connected in a certain way. Sunsets normally produce certain types of experience, and the individual experience-tokens come to represent a sunset in virtue of the fact that they belong to that type. Accordingly, if a particular token-experience is caused in a deviant way, it may still represent a sunset. This happens in cases of hallucination, where there is *mis*representation. Misrepresentation is only possible provided that what gives a state its representational content is how that type of state is normally causally connected. Failing that, a state would always reliably represent its whole causal chain, and there could be no misrepresentation.

I do not hope to show that internalism is wrong and that externalism is right. But I think it is clear that there is a strong case for externalism, and we need to take it seriously. Heil has said, rightly I think, that there is something to be said for working on the assumption that externalism is right, if for no other reason, than that it presents a 'worst case possibility' for reasons as causes. If we can solve the problem of the role of content in causation assuming externalism, the job should be much easier given internalism. For this reason I will proceed on the basis that content supervenes on a broader base than merely the internal states of agents. *Thoughts acquire their content, at least in part, by the causal and historical connections they (and others like them) have with the world.*

But now we face a problem. On the one hand, we have certain common-sense ideas about the causal relevance of reasons. Let us say,

(1) The content of mental states makes a difference to their action-producing character.

On the other hand, we also have a commitment to the view that intrinsic properties alone are causally relevant:

(2) (i) All causal properties of objects/states are intrinsic properties of those objects/states

If externalism is right, we also say,

(2) (ii) The content of mental states is an extrinsic property, at least in part.

Together, 2(i) and 2(ii) imply

(3) Content is not a causal property.

But (1) and (3) are incompatible. We may therefore have to conclude that our pre-philosophical commitments were mistaken. Though it was supposed that reasons cause action, and that the content of our psychological states plays a role in producing action, it now seems as if reasons cannot do causal work, at least not in virtue of their content. If we remain committed to the view that mental processes are responsible for action, it seems that we must suppose, instead, that mental states cause action in virtue of their having certain physical characteristics.

This view is supported by another consideration. We want to explain mental to physical causation within a naturalist framework. This means we must insist that physical effects, such as behaviour, are caused by physical events. Thus, every physical event must have a physical cause. This is known as the 'causal closure of the physical domain'. Were we to reject this position, we would have to face the problem of Cartesian interactionism, and explain just how non-physical causes can have physical effects. Thus far no one has managed to do that satisfactorily, and I will therefore assume that causal closure is correct.

But this yields a further problem. When you have a physical cause for a physical event, irrespective of whether there is a non-physical one as well, it seems as if you have all the antecedent conditions you need to explain the event. A neurophysiological explanation of behaviour (citing only intrinsic properties of states) seems to be entirely sufficient to explain behaviour causally. In light of this, let us reconsider the original example, the happenings at the bar, and adapt a quotation from Norman Malcolm:

> '[T]he movements of [Lofty taking Thirsty's keys] would be completely accounted for in terms of electrical, chemical and mechanical processes in his body. This would

> surely imply that his desire or intention to [stop Thirsty from driving] had nothing to do with his [taking the keys away]. It would imply that on this same occasion he would have [acted] in exactly the same way even if he had no intention to [stop him from driving], or even no intention to [take his keys]. Given the antecedent neurological states of his bodily system together with general laws correlating these states with the contractions of muscles and the movements of limbs, he would have moved as he did regardless of his desire or intention' (Malcolm 1968: 52–53).

If Lofty would have moved as he did irrespective of his desires and intentions, we seem to be in the strange position of having to concede that psychological reasons have nothing to do with the genesis of action, unless we are prepared to assert that non-physical causes can explain physical effects.

> 'We confront a dilemma. Either we concede that "purposive" reason-giving explanations of behaviour have only a pragmatic standing, or we abandon our conception of the physical domain as causally autonomous' (Heil 1993, Preface).

We already know that we cannot settle for the second horn since we are committed to causal closure. But the first horn is not palatable either. It requires us to concede that our common-sense notions are wrong. We would have to accept that reasons-talk, though perhaps convenient, has no deeper basis. On this view, giving reasons to explain behaviour is, at best, a convenience which allows us to avoid giving detailed accounts of physical causes by, instead, 'picking out' an accompanying contentful state as a means of identifying the relevant physical state. At worst, reasons-talk is misleading and should be dispensed with. Either way, the supervenient mental state, since it is not causally efficacious, could be eliminated[17] from the description of the causes, and the explanation would not be the poorer for it. Appropriately, these options in both guises, have been called 'eliminativism'. To wit: eliminate content from causation or causal explanation. That is going to be our last resort. Before we accept defeat we should try every other path. But what is one to do?

4. Two options: sharing labour and dissolving the conflict

We might try to reconcile the horns of the dilemma. We might say that both conclusions can be accepted, without contradiction. We observe that it has not been proven that every cause of a physical event is a physical event. Only that every physical event has a physical cause. The possibility still remains that we can give content work to do, alongside purely intrinsic physical properties. There are two ways of doing this. First, we could argue that mental and physical properties are partial causes, each necessary, but only in combination sufficient for action. Second, they could be independently sufficient for action. Thus, either mental or physical causes might suffice for action. Unfortunately, neither option offers any help. The first violates the closure principle since it requires (albeit partly) a non-physical (mental) cause for a physical effect:

> 'It regards the mental event as a necessary constituent of a full cause of a physical event; thus, on this view, a full causal story of how this physical event occurs must, at least partially, go outside the physical domain' (Kim 1989: 44).

This option therefore does not improve the original position. The second amounts to saying that physical behaviour is causally 'overdetermined'. Had the physical cause not done the job of causing behaviour, the mental would have done it, and vice versa. But the counterfactual case in which the physical cause falls away and the mental cause does the job by itself is unacceptable thanks to the closure principle. And according to that, no non-physical cause can ever account for physical events. But there is a further consideration against this view.[19] If we are committed to the view that the mental is somebow 'dependent' or 'supervenient' on the physical, it is difficult to see how the mental should do causal work independently, or without, a physical cause. Conceiving of the mental doing causal work in such a way takes us back to the problems of Cartesian interactionism. These considerations point to the fact that the notion of causally overdetermined behaviour is not useful for our purposes.

And so the problem remains. Kim calls it the problem of 'causal explanatory exclusion':

> '... a cause, or causal explanation, of an event, when it is regarded as a full, sufficient cause or explanation, appears to exclude other independent purported causes or causal explanations of it' (Kim 1989: 44).

It seems as if we are faced with a very fundamental incompatibility of the two types of solution. Either the mental does causal work, or the physical does it. Reconciling the horns of the dilem-ma is impossible. But naturally, we cannot be expected to accept the position the dilemma puts us in. The only remaining way out is to say that attempting to solve the dilemma is like charging a philosophical windmill. One might argue as follows: 'The alleged incompatibility between physical and mental causes is nothing but illusion. We talk as if mental and physical causes are separate but equal causal forces. In reality, they are not distinct entities. Our linguistic conventions mislead us, and we face a pseudo-problem.'

There are three ways of putting meat on an argument along these lines. First, one might argue that a 'mentalistic' explanation, citing the content of mental states as being causally relevant, is unnecessary. Content can be eliminated in causal explanation. But we have already outlined reasons for avoiding 'eliminativism'. Taking this line will be our last resort. Second, we could argue that mental events are identical to physical events, and that, therefore, we can reduce mental explanations to physical ones. The apparent incompatibility of the two 'types' of explanation will then evaporate: the one type of explanation is identical to the other. And third, we can argue that mental events are causally relevant because they are all physical events, but that the mental remains, irreducibly, a separate domain. We must now examine the two latter options, in turn.

The 'type identity theory', though not so popular nowadays, was once a widely respected approach to accomplishing the reduction of mental explanations to physical ones. Its motivation was to allow that having a mind must make a difference, while accepting the severity of the problem of interactionism. Accordingly, interactionist difficulties are solved by getting the realm of the mental itito the physical. It is held that a mental event, say ME1, is nothing but physical event, PE1, occurring in a person's brain. Though these appear different, they are the same thing, viewed differently:

> 'Just as we may be presented with one and the same phenomenon in two different ways and subsequently discover the identity, so — it has been claimed — we may be presented in two different ways with a mental phenomenon, physically, and (more familiarly) mentally. An analogy would be this: a substance, such as water, may present quite different appearances when looked at with the naked eye and when examined with a microscope, so

> that it will not be obvious that it is one and the same thing that is thus presented. Similarly, it is said that pain may appear in one way to you who are enduring it and in another to the brain specialist examining your grey matter — yet the same thing is being presented. To make sense of these cases of discovered identities we need a distinction between the property denoted by a word and the concept it expresses: we can say that "water" and "H20" denote the same property (the same type) yet do not express the same concept (have the same meaning)' (Mc Ginn 1982: 18).

Identity theorists argue that empirical evidence may allow us, one day, to show that every mental event, say X's feeling a pain, is accompanied by a certain physical event, for example the firing of a certain neuron-bundle, X-117, in X's central nervous system. Evidence permitting, ME1 may then turn out to be PE1, in just the same way that physicists made the empirical discovery that heat and kinetic energy are theoretically identifiable.

Type identity has become unpopular with philosophers because, amongst other reasons, it has appeared as if type identity is incompatible with the notion that mental events may be realized in more than one way (that they are 'multiply realizable'). We do not believe that creatures of a different composition than ours could not possess mentality just because they are composed of different material than we are. For example, it seems possible that ME1 may be realizable not only by persons whose neural bundle X-117 is firing, but also by aliens, whose silicone-platinum transmitters are discharging, or even, in principle, by a 21st century computer's internal processes. Indeed, we seem to have few or no beliefs about the material basis which must underlie mentality. Rather, it seems, we are concerned about a structural, or functional similarity between agents when we attribute similar mental states to them. Minds seem to be realizable by any number of underlying structures. But the mind-body identity thesis is committed to the view that mental properties map 1-1 on to physical properties and, therefore, that a given mental property is realized by only one physical property, say by an agent's neural property N. Suppose we permit that a certain alien has the same mental properties as human beings. Where is the physical similarity, the coincidence of the Martian's having a certain mental property with his possessing property N? Since Martians have no neurons the coincidence is not discoverable. But then the thesis cannot account for our intuition of the possibility of the multiple realizability of mental events.[20]

Establishing that our dilemma is a pseudoproblem seems not to be possible by a reductionist approach, though we will have more to say about it later. Let us therefore examine the non-reductionist option. On this view every mental event[21] is also a physical event. My being in pain involves my undergoing a certain physical event. But a certain type of mental event, like being in pain, is not simply reducible to a specific type of physical event. Identical mental events (my being in pain and the alien's being in pain) may have different physical bases: my pain just is my being in a certain neuro-biological state, martian pain just is his being in a certain silicone-state. Though every type of mental state is 'underwritten' by a physical event, it need not be any particular type of event. Whereas type identity requires that a mental event type must map 1-1 on to a physical event type, token identity permits a 1-many mapping. A mental event type, like being in pain, can be realized by any number of particular physical events, that is by physical event tokens. Now since every type of mental event is realized by a physical event, we can say that every mental event is a physical event, and so, behaviour is caused by a physical event. To this extent token-identity helps us in the way type identity would have, but without running into difficulties with multiple realizability. Indeed, accommodating multiple realizability is token-identity's *forte*.

I have been speaking casually of mental events being 'realized' by physical events, though this relation is far from easy to understand. According to a position first developed by Donald Davidson (see Davidson, 1970) mental properties are said to depend on, and are instantiated by ('supervene' on) physical properties. The position endorses ontological reduction of supervenient properties, but eschews conceptual reduction. Without going into the finer formulations of the thesis or possible objections to it we may state its thesis briefly as follows: 'a predicate p is supervenient on a set of predicates S if and only if p does not distinguish any entities that cannot be distinguished by S' (see 'Thinking Causes'). The supervenience of the mental on the physical on the mental 'might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect' (*Ibid*., footnote 5). If we understand token-identity in terms of the supervenience theory, token identity means that we call any given physical event a mental event in virtue of the events having certain mental properties, like having content. Now, on the naive, common-sense account we said that mental events cause physical behaviour. Problems arose for that simple view. On this new account we still say that mental events cause behaviour, but we add that they do so in virtue of their physical properties. But this addition does not help us much. The difficult question which now arises is almost the same as before, except that instead of asking what role mental events serve, we ask, what purpose do the mental properties serve? And the answer to this question is identical to the earlier one — none. This is the problem of *epiphenomenalism*:

> '[...] The physical event is doing all the causal work in virtue of its physical characteristics; its supervenient mental features are merely going along for the ride' (Heil, M-B. Problem, p. 10).

The following diagram) gives an illustration of the basic idea (the double arrow represents a supervenience relation, the single arrow a causal relation):

Mental Event
⇑
Physical cause --> Behaviour

Externalism does not pose a problem for this picture, even though it may initially look as if externalism makes supervenience more difficult because content depends on relational properties of the physical causes. But all this means is that we must give supervenient properties a wider base:

Mental Event
⇑
Rel. props + Physical cause --> Behaviour

The essential problem is still there: mental properties are going along for the ride. The only difference being that the role of mental properties in causation needs to be solved, because they seemingly cannot add causal power to the 'here and now' causal characteristics of physical properties. And so the non-reductive option is also unsatisfactory.

On the face of it, the idea of 'dissolving' the problem of the two types of explanation by claiming their identity has not proved helpful. If we claim type-type identity, and hope to reduce the mental domain into the physical, we are faced with the problem of multiple realizability. If we avoid reduction by opting for property-dualism, we get an epiphenomenalism of mental properties.

5. Undermining premises — problems with 'syntacticalism'

Despite various attempts at refining our account of how mental to physical causation occurs we have stumbled across one single problem, modified to suit different formulations of how the mental can cause the physical. It was simply this: only intrinsic properties are causally relevant, content is not an intrinsic property, therefore, content is not causally relevant. Extrinsic properties are out of work. But how do we know that intrinsic properties alone are relevant? We assume that in hypothetical, counterfactual cases an agent's behaviour is not affected by the different extrinsic properties his states may possess *provided the intrinsic states remain the same*. Here is an illustrative example (Putnam, 1975): Imagine that there is a world exactly like ours, a molecule for molecule identical 'twin earth', with just one difference: what twin-earthians call 'water' is not H20, it is of different composition, XYZ. When Lofty utters the words, 'Here, Thirsty, I think you'd better have some water', then his thoughts concern water, partly, because there is water on earth. On twin-earth, Lofty's twin's thoughts will concern something else, and the content of his thoughts will be different. We know that people on twin-earth behave in just the same way as they do here. Why, then, is the behaviour identical, given that the intrinsic properties of the internal states of Lofty and Lofty^1 are identical, and given that extrinsic properties are not shared? Mill's Method of Agreement and Difference confirms the common-sense view that the causally relevant properties of systems which behaved identically are those they shared — the properties of their intrinsic states. Despite extrinsic differences the systems act in the same way. Hence extrinsic differences seem to be causally irrelevant.

This is the conclusion of the argument from 'syntacticalism', according to which only the syntactical, formal (i.e. intrinsic) properties of internal states are causally relevant. The problem of the apparent causal-explanatory exclusion of mental properties arises as a result of the fact that we accept syntacticalism's premise, that causally relevant properties are what a system and its twin share. It is supposed that whatever two systems which act identically do not share cannot be relevant to their behaviour, on the basis that, apparently, the behaviour can occur in the absence (in the *other* system) of these factors. This conclusion I will call Mill's Principle.

Some reflection will reveal that this principle is not uncontentious. It appears as if in certain circumstances the principle can lead to counter-intuitive conclusions. Imagine a flourishing tulip which feeds only on sun and water. Sunlight permits life-sustaining photosynthesis. Suppose further that this tulip has a doppelgänger-tulip on twin-earth: tulip. But tulip feeds on sunlight and 'twater', a substance of chemical composition XYZ. Both plants behave identically. Suppose that the shared intrinsic properties are that both plants are photosynthesizing thanks to the presence of sun. But we know that water and twater are not identical in their composition, and so the presence of water in the one plant and twater in the other seems not to be a shared property. Using identical reasoning as previously we ought to conclude that water and twater are not causally relevant to the growth of the plants. And yet it seems evident that what made the one tulip grow was twater and what made the other grow was water.[22]

Clearly, the principle of reasoning which underlies both the tulip and tulip^1 example and the Lofty and Lofty^1 example is that causally relevant properties are those properties which are shared by systems across possible worlds. Does the tulip example prove that Mill's Principle is too strong? We must, I think, allow the insight that the conclusion of syntacticalism does not follow from the premises with absolute certainty. We certainly can entertain the thought that extrinsic properties are causally relevant, and the argument from syntacticalism is not going to disprove that. But it can be doubted that the counter-example proves more than that. It not so much undermines our intuitions about the causal relevance of intrinsic properties as it shows that we may not have been clear about what an intrinsic property is. We learn that not only shared compositional properties are causally relevant. The example elicits intuitions that it does not matter to the tulip whether it gets water or twater, which are not of the same composition. What matters is that water and twater share a structural property. And so it becomes clear that it is often difficult to determine what property really is shared by a system and it's twin, because we may not be sensitive to all the relevant facts. Tulip and tulip^1 do presumably flourish because of a shared internal, structural property. The capillaries of both plants may be of identical size. This property would permit them to utilize any liquid with a certain micro-structure. In other words, the property they both possess enables flourishing irrespective of whether water or twater is in the environment. We can still say that it is in virtue of this shared internal property that the two tulips behave identically. Then the counter-example does not successfully undermine Mill's principle. We simply make it clear that what we mean to say is: *causally relevant properties are those structural properties which a system and its twin share*.[23] Causally relevant properties are intrinsic, functional properties. The attempt to undermine syntacticalism's conclusion fails.

And yet we seem to be unable to get rid of reason-giving explanations. There are cases when physical explanations will not do. Consider the example of Lofty taking Thirsty's car keys away. We are familiar with the reasons-explanation, in which the content of mental states came to bear on action. One alternative is the neuro-biological, physical explanation, which will look awfully daunting: 'Lofty's neuron bundle X-117 of the cerebral cortex fired at frequency F, resulting in synapse transmission of kind S ...' and so on. Does this answer the question why he took the keys away? Not at all. We do not care what happened inside Lofty's body, triggering his actions, whether it was bundle X-117 or bundle X-118 that fired. That might be interesting if we wanted to know *how* it came about that Lofty took the keys. But as this is not what we were interested in. The neuro-biological answer seems to be the answer to the wrong question.

This is a great puzzle. Suppose that the neuro-biological answer is, in its own right, absolutely correct. Why does it not explain what we want explained? Is it not paradigmatic of a satisfactory explanation in view of the fact that it offers antecedent sufficient conditions for behaviour? What more could one ask when requesting an explanation? In order to answer these questions we take a brief look into the nature of explanation.

6. The nature of explanation

In order to see why at certain times some explanations are better or worse than others, we must consider what the demand for the explanation of an event is and how it can be satisfied. I think we can go a long way towards doing that by considering David Lewis' thesis (Lewis, 1986)[24]: to explain an event is to provide some information about its causal history. Note that this does not mean: provide *all* the information about its causal history. This is rarely, if ever, possible. Consider the case of explaining why a lightbulb went on. One explanation might be: 'Someone switched on the switch.' Notably absent from the explanation are such facts as the existence of wiring, connecting the switch and the bulb, or the fact that the power-station is supplying the house with power. Or, that two decades ago, a bill was passed in parliament, permitting the construction of a power station. Or, that a responsible lad, who became Minister of Energy, came up with the idea. Yet these are all partial causes of the event. And as for each of the partial causes, these, too, have their causal histories. The whole causal history is thus a very long affair, full of branches and divisions, each with their own histories. As Lewis remarks, 'roughly speaking, a causal history has the structure of a tree' (Lewis 1986: 215). We can only imagine how big the whole cause is. However big that is, it is, in most cases, far too big for us to know in detail.

We sometimes speak as if there is *the* cause, a single definitive cause of an event. But we can never specify the whole cause. What we give is what we take to be the most salient aspect of the whole cause. But in providing a partial cause we do not suppose that it, alone, is relevant. Suppose two people give different answers to the question why the lightbulb is burning. The first explains that the switch is turned on, the second says that the bulb-filament is still intact. Both would agree that the other is providing an explanation. Which is the better one? This depends on the situation. If the interrogator lives in a house in which bulbs are always blowing, the second explanation will be quite sensible. And if the person is leaning against the switch without knowing it and asks the same question, the first will be better.

If we always fall short of giving the whole causal history we must be selective in giving a partial one. We are not only bound to be selective about which part of the history we give, we also have to decide which aspects of that section we give. For even there we cannot hope to give all the relevant facts. Take the explanation which cites the fact that the filament is in order. This, in a sense, is the part of the causal history which is being focused on. But that section is far from fully described. What is the length of the filament? Were it 10 nm long, we would not be able to see the light. Is the filament in a vacuum? These are just some of the endless questions to which a full answer would need to give answers.

Would the fact that a certain causal history is more detailed than others mean that it is better? In other words, is there an ideal explanation (the whole causal history?) which we must come as close to as possible, and the closer we approach it, the better? Consider another example. A social worker notices that a relation of constant covariation exists between levels of pollution in drinking water and the manifestation of cholera in the drinkers of the water. 'Pollution causes cholera' is her conclusion. Suppose that, in fact, a little organism causes cholera. The organism thrives in polluted waters. So pollution only indirectly causes cholera. Has the social worker made a mistaken inference, given a bad explanation? No. Pollution does cause cholera, in virtue of its causing the relevant organisms to thrive. But the explanation is incomplete. Suppose, however, that the social worker is not interested in biological details. Then there is no reason to prefer the fuller explanation. Unless, of course, it turns out that the predictive powers of the theory are extremely important. Perhaps the fuller explanation is more sensitive to changes in the environment that would affect the growth of the organisms. Suppose that in soil which is very acidic the drinking water cannot sustain these organisms. Then the fuller picture is better because it, alone, can explain the absence of cholera in communities living on acidic soil. But notice that this superiority is dependent on the interests of the researcher, in this case predictive reliability. There are cases when the interests of the inquirer do not necessitate a more complex history. If the researcher is not scientifically educated, for example, she may well feel insecure when the explanation for the phenomenon she is interested in demands a biological explanation. Further, cholera may, in fact, covary very consistently with polluted drinking water. Then her explanation seems better than the fuller one, because for the purposes of preventing cholera in future (which is her aim) the two are comparable, whilst the simpler explanation (in contrast to the fuller one) will not cause her to lose confidence in herself when recommending preventative measures to the community.

We must resist the temptation to say that if the researcher does not know about the organisms, she cannot explain the incidence of cholera. Such a view amounts to a prejudice in favour of an explanation most interesting to us. Even though there is more to the causation of cholera than the researcher has acknowledged we cannot deny that she knows quite a lot about the causes of cholera.[25] Therefore, giving partial histories, as against whole ones, does not preclude us from giving explanations. This is a fortunate conclusion, since if it were any other way we could not explain anything.

On this view, there is no definite standard of explanation, no standard 'serving'. The basic rule is, the right amount of information is relative to the amount you want. Lewis makes a nice analogy:

> 'Hempel writes: "To the extent that a statement of individual causation leaves the relevant antecedent conditions, and thus also the requisite explanatory laws, indefinite, it is like a note saying that there is a treasure hidden somewhere." The note will help you find the treasure provided you go on working. but so long as you have only the note you have no treasure at all; and if you find the treasure you will find it all at once. I say it is not like that. A shipwreck has spread the treasure over the bottom of the sea and you will never find it all. Every dubloon you find is one more dubloon in your pocket, and it is also a clue to where the next dubloon may be. You may or may not want to look for them, depending on how many you have so far, and on how much you want to be rich' (Lewis 1986: 237).

Whatever you need explained, you will always get a partial answer. How much information you get will depend on how much is known and how much you want. The latter is usually fairly readily discoverable. In Lewis' words, 'When partial answers are the order of the day, questioners have their ways of indicating how much they want.' There are several ways of doing this. Interrogators may use 'contrastive why-questions'. For example, they may ask, 'why did Lofty take away Thirsty's keys, rather than knocking him out with a punch on the head?' The answer will not include information which would have been appropriate to answer the contrastive question. Thus the scope of suitable explanatory information is reduced. Contrastive stressing has the same effect. 'Why did Lofty take the *keys* away?' also supplies us with an implied contrast: '... and not his shoes or his hat or his watch?' Again, the scope of suitable information in the explanation is thereby limited, making it easier to decide which information is relevant.

I believe there is a third and very important way in which the question limits the scope of appropriate answers. The words in the question, even when neither stressed, or used contrastively, indicate the kind of detail required. On the one hand, there are various specialist discourses, and if the question is phrased in the discourse of one such discipline, then an answer in the discourse of ordinary language is likely to be inappropriate. The converse also holds. If the question is phrased in ordinary language a specialist response will be no good, for obvious reasons. The general rule is that whenever a question is phrased in a certain discourse, the response should be in the same. On the other hand, even if we stay within a certain discourse we follow certain rules. Roughly, we must remain in the right register, address the relevant concerns and at the tight level of generality. Suppose I ask a question in which the liquidity of water is relevant, like, 'why do steel boats float?'.[26] The words used in this question are all 'household items'. This limits the degree of specificity desired in the answered. For example, I cannot get into micro-physics and the forces of attraction and repulsion between molecules in liquids. The appropriate 'level' of language will go as far as mentioning buoyancy, pressure, perhaps even density, and liquidity. The advantages of doing this are not merely that the questioner will better understand the answer. It is likely that the question was motivated by puzzlement due to an observation which contradicts a previously held belief that heavier-than-water objects ought to sink. This is confirmed by the presence of the word 'steel', which is perhaps tacitly contrasted with 'wood'. Though the micro-physical explanation will explain why that steel floats in that water, it will not explain why items constructed from solids heavier than the supporting liquid medium can float. For example, it will not explain why iron-wood carved in a certain way floats in mango-juice. Thus the explanation is at a level which is too specific, and the explanation cannot be applied to the appropriate level unless the recipient of it can translate the terms back into ordinary language.

Besides the right degree of detail and generality, there is another important requirement for an explanation. It must possess the right perspective on the event. An explanandum can be described in a great number of ways. The killing of a president can be described as the death of a person. Or as the discontinuation of bloodsupply to the brain of a mammal. But it can also be described as the death of a husband, or the death of a statesman. Or, it may be the trigger for a revolution, or the end of an era of peace. In short, it matters a great deal which aspect of an event you focus on. If the question 'views' an event one way, the explanation must address the same position. 'Why was the killing of the president a bad thing?' We cannot answer, 'Because his brain stopped getting blood.' And so the question supplies us with clues about how the explanandum event is to be described.

Nothing I have said is supposed to constitute a hard-and-fast rule about explanations. But there is general applicability in these observations, which is enough to give us an insight into how the scope of possible information given in explanations is limited. And we are now in a better position to answer the question posed at the end of the last section: When asking for an explanation of an event, what more could you want than its antecedent conditions? It is clear that an explanation *need* not offer antecedent necessary or sufficient conditions at all. Sometimes, antecedent conditions that were merely involved in the causal chain are enough. Furthermore, we need to be given the right part of the whole causal history, and that part must cover the right degree of detail, so that the explanation is at the right level of generality. And the same aspect of the event must be explained as was inquired about.

7. Appropriate responses

Let us now get back to the question why the neuro-biological explanation does not answer the question, 'Why did Lofty take Thirsty's keys away?' The problem is not that the neuro-biological answer fails to account for the specified event's coming about. An important part of the causal history has been specified in detail. But only a specific physical event has been explained, in other words, only a particular sequence of movements. Had Lofty taken the keys away by asking the waitress to take them out of Thirsty's pocket, the original physical explanation would no longer apply to that event. And yet, in the sense meant by the questioner it would be the same event. The specifics of how exactly the process of key-taking happened are presumably of no interest to the interrogator. What matters is why a customer was relieved of his means of getting home. And an explanation must explain *that*. The neuro-biological explanation is a causal history of a specific event, which is only one of many possible instances of a customer being denied the right to drive. The neuro-biological story is designed to explain that specific event, not as an instance of 'stopping a drunk customer from driving', but as a sequence of bodily movements. The discourse of neurobiology explains this sequence by referring to certain neuron bundles, say X-117, which figure in bringing about one movement-sequence. As we saw in the last section, there are many rules of communication which determine what information that explanation gives us. In the discourse of neurobiology the fact that bundle X-117 fired probably has certain implications, for example, that X-118 did not fire. The fact that it fired at frequency F means it did not fire at frequency E. And these differences matter in the discourse of neuro-biology.[27] We can imagine that for a neuro-biologist, had bundle X-118 fired at E, the explanandum event would have to be regarded as an anomaly. However, the given statement of these facts does not serve to explain the event in the way we want it to be explained. The right explanation would cover other possible instances of key-taking, for example, with the left hand, with the help of a waitress, and others. And it will do that by using an explanation which generates tacit contrasts with events which are, loosely, at the same 'event level'. This level is far above the discourse of neurobiology. It is the level of ordinary language which is where we identify events as actions.

All this seems to indicate that questions phrased in the language of a certain level cannot be replaced by questions phrased in the language of another level. However, this would be a premature conclusion. Though for pragmatic reasons this does not often happen, there is nothing to prevent it in principle. Here is a case in point. Let us regard the word 'runnings' as denoting a class of individual activities, all of which could correctly be described as 'runnings'. These individual events are each members of the higher-order class 'runnings'. They are: 'jogging', 'sprinting', 'bolting', 'fast bipedal locomotion', and so on. In principle, it is possible to list all members. Even if we cannot do it, God could. Then a possibility arises. The class of 'runnings' can be specified in more than one way. First, obviously we could refer to the class as 'runnings'. But second, we could refer to the same collection of runnings by listing the members of the class, collectively. There is nothing more to the class than the collection of its members.

Suppose that we are interested in the question whether running is dangerous in panic-situations. In the explanation we cannot replace 'running' with a member of the class of runnings, even though every class member is an instance of running. For example, we cannot offer an explanation of why jogging is bad in such situations. This would leave open the possibility that, in contrast, sprinting is all right. Thus, in an explanation we cannot substitute a class with a class-member. However, suppose we explain that running, jogging and all the other members of the class 'running' are bad because when people see others moving in those ways they also start running, which causes jams at the doors. This explanation is as good as one which just refers to 'runnings'. Since 'runnings' designates the same collection of activities as the above-mentioned specification, the terms are replaceable. Similar examples could be created for other words. The basic rule would be the same: to replace a higher order class term, substitute it with the list of its total membership.

How is this of interest to us for our project of trying to account for the differences between physicalist and intentional explanations? We have here a case where language of one level is reducible to language of another. One of the factors preventing us from replacing intentional language with physical language was that physical explanations explained too narrow an aspect of what needed to be explained. But here we have a similar case: explanations citing the individual members of the class of 'runnings' are also much more specific than those that mention the higher order class. And yet, here we managed to replace specifications of the higher order kind with member-lists. This raises the possibility that we could do the same with intentional language in explanations of action. And the first step towards testing that possibility is to re-examine the question whether mental events cannot be reduced to physical events.

8. Reducing the mental

Recall that when we tried to claim type identity between a mental event and a physical event we encountered difficulties. We found that a mental event type is realizable by any number of physical event tokens. It seemed as if this one-many relation ruled out reduction. But the considerations raised in the last section indicate that a one-many relation between a class and its members does not prevent us from reducing a class to its total membership. We may now be able to rewrite story about mental reduction soniewhat. Using the language of supervenience, we might say the following:

> 'Imagine that we discovered that I realize a thought of Vienna in virtue of my possession of neural characteristic *N*. Suppose that an Alpha Centurian realises the thought in virtue of his possession of O, and that a computing machine realises the thought in virtue of its possession of P. We can still say that each of us realises a thought of Vienna in virtue of our possession of a single property, namely the disjunctive property, *N* or *O* or *P*' (Heil, M-B. Problem: 13–14).

'Thinking of Vienna' is a mental event which supervenes on processes with the 'higher order' (disjunctive) property of being *N* or *O* or *P*. If we suppose that these are the only ways in which a thought of Vienna may be realized, the difficulty with multiple realisability no longer stands in the way of our saying that thinking of Vienna just is a physical event with the property *N* or *O* or *P*. Thus, my thinking of Vienna has no properties which is not also shared by an event with physical properties *N* or *O* or *P*. The introduction of 'higher order' properties may enable us to reduce mental events to physical ones, this time utilizing a reduction base wide enough to permit a one-many mapping onto lower-order properties (like *N*, or *O*, or *P*). The higher order property 'being *N* or *O* or *P*' is realized by any one of the lower order properties. Heil anticipates that one may have doubts about this manoeuvre:

> 'Is this cheating? Not at all. It is important to bear in mind that N, O and P designate, not properties, but characteristics of complex objects that happen to be salient to us. For all we know, the predicate "N or 0 or P" might designate an important natural kind, a kind that, owing to our taxonoffic practises, appears ad hoc or gerrymandered, but is, from the point of view of the natural world, perfectly simple' (Heil, M-B. Problem: 14).

There is nothing mysterious about a predicate like '*N* or *O* or *P*' because it does not in fact designate a disjunctive property. A disjunctive predicate of 'being *N* or *O* or *P*' must be distinguished from a disjunctive property like 'being red or not being red'. That would be a gerrymandered property, worth shaving off, provided we have a Russelian 'robust sense of reality'. But Heil's predicates are something entirely different. They are simply disjunctive specifications of what may well be unitary properties.[28] It is indeed hoped that objects possessing a common, disjunctively specified property share a relevant intrinsic unitary property, of a perfectly natural sort. Hydrochloric acid and sulphuric acid both posses the 'gerrymandered-looking' property of 'being able to dissolve cotton'. But this property in fact supervenes on a shared 'natural' property, since both acids have certain chemical structures which cause the dissolution of cotton. Then we must not dismiss the higher order 'property' as being artificial.

Some reflection reveals this to be plausible. What, for example, do bananas, soap and oil have in common? They are all slippery. Though slipperiness is a dispositional property, that disposition depends on, or is realized by, certain intrinsic properties of all three. These intrinsic properties are shared, even though the composition of the three is different. Some structural property is shared, such as weak inter-molecular bonds, or whatever. The extrinsically specified shared property 'slipperiness' is thus realized by an internal shared property. There is nothing strange about 'getting at' an intrinsic property by means of an extrinsic, dispositional description of it.[29] In many cases, we may often only come to suspect an intrinsic similarity one we have seen the extrinsic similarity. For example, imagine that we get our drinking water from three sources. Initially, the only thing we know about the water is that in two cases, the water has the (extrinsic) property of causing cholera. As microbiology can show, there is also a natural, intrinsic property, the presence of creatures, which explains the shared extrinsic one. Similarities in events or objects are initially often only discoverable at the level of 'gerrymandered properties'. And this means is that Heil's 'disjunctive properties' have something to be said for them.

The distinction between properties and specifications of properties yields a further suggestion. I have tried to show that specifying a property disjunctively is equivalent to specifying it through a mental description. Now one could plausibly suggest that a mental property just is a physical property, which has simply been mentally specified. In other words, when we call a property a mental property (like we call the property of having content a mental property) we may do so because we cannot put our finger on which physical property, exactly, is involved. However, we do not doubt that some physical property is involved —neither if we believe that the mental supervenes on the physical, nor if we believe that the mental is reducible to the physical.[30] This being so, we have to specify the physical property in an indefinite, and perhaps therefore — *mental* way. When we specify a property mentally we are actually hinting at a disjunctively specified physical property. Then, when we reduce the mental to the physical, what is actually being reduced is mental talk.[31]

Jerry Fodor (1975) argued that reductionism of the sort discussed in this section is probably a mistake. One of his claims is worth mentioning because it shows just how essential the above-mentioned distinction between kinds and specifications of kinds is. Fodor says:

> 'If it turns out that the functional decomposition of the nervous system corresponds precisely to its neurological decomposition, then there are only epistemological reasons for studying the former instead of the latter. But suppose there is no such correspondence? Suppose the functional organisation of the nervous system cross-cuts its neurological organisation. Then the existence of psychology depends not on the fact that neurons are so depressingly small, but rather on the *fact that neurology does not posit the kinds that psychology requires*. I am suggesting roughly, that there are special sciences not because of the nature of our epistemic relation to the world, but *on the way the world is put together*: not all the kinds (not all the classes of things and events about which there are important, counterfactual supporting generalisations to make) are, or correspond to, physical kinds' [my ital] (Fodor 1975: 24).

Fodor seems to think that when we use the taxonomy of a given discourse we are committing ourselves to the existence of certain kinds, which the special discourse is most able to 'pick out'. But he is being 'realist' about kinds of the special sciences in a way we need not be. We may specify a kind non-physically, as we do whenever we use mentalistic language, without thereby being committed to the view that we are actually 'picking out' non-physical kinds. We can have the benefit of describing the world in the way we need to, thereby permitting us to detect patterns and make predictions of the right sort, without making any commitments about the true nature of the underlying kinds. Indeed, we may be very reluctant to grant the distinction between physical and non-physical kinds in the first place. After all who knows what kinds there really are? We do not have access to them, and there is no need to make unnecessary assumptions about them so long as we can use different taxonomic practises. And this we can do without shaky assumptions about 'how the world is put together'.

This attempt to show that the mental is ontologically reducible to the physical appears to be more successful than the last. This is good news since we now have reason to hope that intentional language in explanations of action should in principle be replaceable with physical language, with consideration for requirements arising from the rules of communication of the sort discussed in Section 6. The old hopeful thought arises: if we are genuinely committed to working within a naturalist framework, how could intentionality involve anything *but* physical objects and events? If we are really committed to the view that physical events must have physical causes, then we had better account for reasons and contentful states in terms of natural phenomena. However, the harsh fact is that we do not use intentional and physical language interchangeably. What is the difference between intentional and non-intentional systems that necessitates this? To answer this we move on to discover the nature of intentionality. One might say that our problem as naturalists is to find out how to put together the physical building-stones in order to build an intentional house.

9. Building intentionality

If intentionality is to be realized in physical systems we are faced with a construction problem. How can we build a machine which has reasons for acting? And what would such a machine need to have 'built in'? In this section I want to propose some of the crude requirements needed to 'construct' intentionality.

Not any physical construction can realize intentional states. Take a thermostat as an example. Bimetallic strip variations covary with temperature changes, and two causes circuit closings and openings, which control the air-conditioner. This machine is so simple that it is certainly wrong to argue that,

> '... machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines of problem solving performance' (McCarthy, quoted in Searle, 1991: 513)

Two problems leap to mind. First, there is the question of what sorts of beliefs a thermostat might have. Hardly these: 'It's too cold in here' and 'It's too warm in here.' Thermostats just do not have the complexity of construction or of behaviour to manifest beliefs. And there is general agreement in the view that beliefs come in systems, not in singles or in pairs. Second, it is difficult to see which problem the thermostat is trying to solve. If it is the problem of getting the right temperature, we must ask, whose problem is that? When we ask, 'Why did the thermostat switch on the airconditioner?' we explain the behaviour with reference to some person's aims, like, 'The engineer designed it to keep the room temperature at 22 degrees'. And in cases where we talk as if the machine had reasons for actions, we simply metaphorically 'extend our own intentionality' (see Searle, 1991: 511).

The thermostat is merely an 'organ' in the physical chain of events leading to the satisfaction of a person's desires or design problem. And as we saw in Section 2, in cases of intentional action the reason why the system acted must be its reason. Aristotle said, 'the origin of action is choice, and that of choice is desire and reasoning with a view to an end'.

The thermostat does not have ends, or reason towards them, and this inclines us to say that it cannot act, that it is not an intentional system. I think we can accept the view that, at least, intentionality involves a capacity to reason, and a capacity to have aims, or goals. Perhaps we attribute goals to systems automatically, provided we attribute them with an ability to reason first. Let us take this for granted. Then we must establish when we attribute an ability to reason.

It is possible to come up with designs which are so complicated that they come much closer to reasoning than a thermostat. Sophisticated computers almost seem capable of it. And yet they don't quite seem to get there.[32] As John Searle has noted (see Searle, 1991), they simply manipulate formal symbols without regard for their meaning. In other words, whatever computers do, they do because of the physical properties of the processing-states. And the meaning of those sates is irrelevant to their action. In this sense, computers and other non-intentional systems are 'blind executors of designer wills'.

In line with the view expressed by Searle, Fred Dretske (1990) has argued that an intentional system must detect representational properties of internal states and utilize them in the choice of output. This means that the content of representational states has a role to play in the production of bebaviour. Contrast this with the thermostat's case. Let us grant that the bimetallic strip position represents room temperature.[33] The thermostat will close and open the circuit irrespective of what the bimetallic strip represents. Thus, only the formal, syntactical properties of the strip matter to the thermostat's behaviour.

Dretske and Searle present us with a useful distinction regarding a system's ability to reason. It may be presented as follows: 'Reasoning and processing are not the same thing. Processing is just a protracted process in which the formal, intrinsic properties of system-states result in further states. In short, processing is just a blind, physical occurrence. But reasoning is not blind. Here the representational content of internal states is taken into account, and further states result not in virtue of the physical properties of preceding states, but in virtue of representational properties. The system "sees" the content of its states.' This view accords well enough with our common-sense ideas. But is there not something deeply mysterious about the requirement that a system should behave as it does because of the semantic properties of its internal states, and not their intrinsic properties? After all, if the system is a physical system it seems very likely that everything it does is caused physically. Searle suggests the right line of thought:

> 'Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena' (Ibid., p. 519).

So clearly, in one sense a system does do what it does because of the intrinsic properties of its internal states. But then, in what sense must we take the view that a system behaves as it does because of the semantic properties of its internal states? Notice that our problem very closely resembles the one we were originally faced with: if a physical explanation will do the trick, why should we also offer a mentalistic one? In the present case, the problem is, if formal properties are sufficient to cause behaviour, why do we need semantic properties as well? The solution to these parallel cases will turn out to be closely linked, too. But for the moment we accept defeat: we did not answer the question what is needed for intelligent reasoning. And so we face the issue of intentionality again, this time directly.

It may help us if we imagine an intentional machine at work. We might imagine the machine as a robotic apparatus which can move about in all terrains and keep from falling off edges and crashing into things. It has cameras built in, as well as audio-equipment which enables it to avoid unforeseen accidents. This machine is also able to find vegetables and fruits by following their smells, and by image-matching what the camera shows of the world with preprogrammed instructions. It is able to ingest the food it finds, and convert the carbohydrates into ethanol, to keep the generator going.[34] This machine, in short, is extremely self-sufficient, and functions remarkably well. Though the machine was designed to behave in a certain way, this must not dispose us towards saying that it has no intentions of its own. After all, in some sense, human beings are also 'designed'. If our instincts balk at the idea of giving such a primitive machine intentionality we simply design one which is more complex, one which replicates the behaviour of small amphibians. And failing that, one that replicates little mammals like rats. All this in order that we fell comfortable about saying that the system reasons, and that it has goals and needs. Furthermore, this machine exhibits very complex behaviour, and it responds effectively to a wide range of situations. The range of situations it can respond to, and the range of responses it has, is so wide that it is almost impossible to give a full description. In short, let us take it for granted that we are dealing with an intentional machine.

We observe this machine for several minutes. In this time it ingests and utilizes food twice. Suppose we know how the machine works because we designed it. The first time the machine ingests a strawberry. This food-getting is realized by a certain causal sequence, abbreviated as C1. The second time a different sequence occurs, C2, and the machine ingests a cabbage. Cl and C2 realize, in a manner of speaking 'food-getting behaviour'.

Now we ask ourselves, what is the difference between these three questions:

(1) Why did the thermostat close the circuit?
(2) Why did the machine eat a strawberry?
(3) Why did Lofty take the keys away from Thirsty?

Here are some important differences. In (1) a physical explanation will satisfy the interrogator. In (2) a physical explanation will probably satisfy the interrogator. We might respond, 'well, it went into state Cl'. This causal history represents the machine's ingesting the strawberry. But, notice that there is also room for an intentional explanation. 'The machine needs food to stay functional.' And in (3) the neuro-physical explanation will be totally unsatisfactory. Only the reasons-explanation will answer the question.

How do we explain the differences, given that in each case a physical system operates, bound by the laws of physics and acting out what it was designed to do? I think we must recognize that there is not clear-cut line between cases when we do ascribe intentionality and when we do not. However, when we do, it is for very good reasons. Let us re-examine question (2). Suppose that the questioner asks it like this: 'Why did the machine eat that strawberry?' The implied question is, 'in contrast to rolling over it, or ignoring it?' Our ability to give a 'technical' explanation depends on the fact that we know exactly what causal sequences underlie the behaviour. We can say, 'C1 realizes strawberry-getting behaviour, in contrast to C2, which realizes cabbage-getting behaviour, in contrast to G2 which realizes rolling-over-the strawberry-behaviour.' These are satisfactory technical explanations. But now suppose we do not know the details of the causal technical histories, because the machine is too complicated and we have not studied it well enough. Attempting to give a technical explanation, the best we could do is, 'the machine was in a state which realised strawberry-getting behaviour.' But this is a useless explanation, since it is trivially obvious. It is like explaining, 'aspirin relieves pain because of its analgesic powers'. Given our technical ignorance, the best we can do is to say, 'the machine realises that there is a strawberry, and it wants to ingest it.' This is an intentional explanation. And it appears to be perfectly satisfactory. This indicates that intentional explanations take over when technical, detailed causal histories are either impossible, or extremely difficult to give.

It is then hardly surprising that we do not give a technical explanation in the case of human action. It would be very awkward indeed, sometimes impossible. However, there is another reason why we do not. Even if we knew the states which underlie a certain instance of human action, unless we also know the states which underlie all other possible instances of the same kind of action, me must use the intentional explanation. Remember that when discussing a possible reduction of mental properties to physical properties we had to use a disjunctively specified physical property as a reduction base. Being painful just is being in an N-state (in humans) or being in an M-state (in Martians) or being in a P-state (in a computer). But if we do not have the full list of appropriate disjuncts, we are in trouble. Suppose Plutonians realize mental states when in a T-state. Then pain is not just the disjunct given above. The T-state is missing. If we want an accurate reduction, and failing god-like omniscience, we must say something like, 'being painful just is a list of disjuncts, whatever they may be, including states N, M and P'. Imagine that now I am in pain. What is the use of saying that I'm in a certain state with a disjunctive property, including such-and-such disjuncts, and including unknown disjuncts? There is no gain in this, and furthermore, it is cumbersome. So we cannot avoid using intentional language unless we can give the full reduction base of an action, that is, in all its manifestations as physical occurrences.

The situation is similar for the attribution of intentional attitudes like desires and beliefs. Ordinarily we say, for example, 'Lofty desired to save his car.' If we convert this to property-talk, we get, 'Lofty was in a state with the property of realising a desire to save his car.' This is clumsy, but correct. How can we reduce such talk? Suppose that the domain of possibilities was very small that day, and there were exactly three possible ways in which Lofty's states might have realized that property: S1, S2 and S3. Suppose also that no other person or thing in the universe might have realized that state (as unlikely as it is), and that we are omniscient, and we know that these are the only ways in which they could be realized. Then we may reduce as follows: Lofty's being in a state with the property of realizing a car-saving desire just is his being in a state with the disjunctive property, S1 or S2 or S3. If the Heilean disjunctive properties work, as we said they do, this will be correct. The only rub is that we are not omniscient. But this thwarts our reduction plans, and at best, we can offer an indefinitely specified reduction base, as suggested in the above paragraph. But as we saw there, this is less convenient than using intentional language, and it serves no better purpose. The very point of using intentional language is to permit the classification of an event in terms of relevant properties shared by a vast and indefinite reduction base. The classification, the intentional-language term, assimilates all the disjuncts in a schematic way. This saves time, makes life easier, but most of all, it ensures that we identify properties at the right level of generality. This means, for example, that we identify an event in virtue of its tendency to bring about such-and-such behaviour, irrespective of what the exact reduction-base of such an event is.

Then we have the following picture. Intentional language which cites psychological reasons in explanations of actions, is, in principle, reducible to physical language. That means, ontologically reduction is in order. But, since we do not know which states underlie actions, we must remain on the intentional level to prevent illegitimate reduction, onto incorrect or insufficiently wide reduction bases. Epistemological considerations thus prevent an elimination of intentional language. The ineliminabilty of intentional language is not merely a matter of convenience — it is unavoidable.

The question raised at the end of Section 7 is thus answered: intentional language can in principle be 'translated' into physical language. And just as we found that mental properties just are disjunctively specified physical properties, so intentional properties turn out to be nothing else than disjunctively specified physical properties. Which means, when we give beliefs and desires as reasons for action, we are in fact giving indefinitely specified physical causes, the specification of which could be made complete if we knew the full reduction base.

Now we can return to the question posed earlier: Is there something mysterious about the requirement that a system capable of reasoning should behave as it does in virtue of the semantical properties of its internal states, and not their intrinsic properties? There is nothing mysterious about it. When we say that internal states are important in virtue of this or that property, we are indicating that there are aspects of these states which are more important than others.[35] In view of this, we must delineate what is relevant to the explanation. We are saying, 'For the purposes of what you what to know, this or that property is relevant.' This is how we must understand the claim that with rational creatures, semantic properties matter.

When we identify an internal state's semantic properties as being the most relevant, we reveal our interest in the explanation of the agent's actions, beliefs, desires, thoughts, and the like. Consider what a neuro-biologist would be interested in: mainly intrinsic properties. Why are we different? We are focusing on rational creatures. When we mention internal states we still want explanations which account for the things rational creatures do, performing actions, and not specific instances of movements which amoebas also realize. And therefore, we are interested in internal states only insofar as they represent something, or lead to certain behaviour, or realize certain actions. We are only interested in this aspect of their states. When we refer to their internal states to explain something, we do so with an explanation of intentional behaviour in mind. This being so, we specify internal states in terms of their relational properties.

Consider an example. A man sees a glass of water and goes to get it, realizing water-recognition states and water-getting behaviour. Let us say that the state in his mind which realizes his recognition is R1. R1 causes an instance of water-getting behaviour, B1. We have here a case that parallels the thermostat cases. Bimetallic strip variation causes circuit closing. Nothing 'rational' about it at all. But here comes the difference: the man is a very complex construction, He can get water in a number of ways. Instead of going to drink the water by taking three steps, our man can take two hops, one leap, seventeen little steps. Or he could lassoo the bottle, or ask a waitress to help. So in fact, S1 is just one of many water-getting states. And what about the water representing state? Not only S1 represents water. There are probably hundreds of states in one human being which represent water.[36] If we must give an explanation of this man's behaviour which focuses on his internal states,[37] we must focus on those properties of the states which explain the sort of thing a common 'intentional' explanation would, so that his getting water is explained. We can see now why we regard rational creatures' internal states as important in virtue of extrinsic properties. But to a certain extent, the very fact that we try to detect differences between intentional and non-intentional systems at this locus may reveal an ignorance of the truly relevant issues. To focus on internal states in cases where actioin are to be explained is to like requiring somebody to explain why running is bad in panic-situations whilst referring only to 'jogging', 'sprinting', 'bolting', etc. It can be done, but only by specifying the relevant properties in a most awkward way, namely, as being instances of the higher order class of 'runnings'. Similarly, to focus on internal states which are normally most suitable for neuro-biological explanations means that we have to specify them in terms of properties which look highly 'gerrymandered'. In short, we have to describe internal states in a strange way when we are using them to explain features which they are not well suited to explain.

In this section I hope to have sketched a picture of how intentionality can be accounted for within the natural domain. To summarize: we do not need intentional language when we explain the functioning of thermostats and computers, we do when we explain the actions of complex, intentional systems. But they are made from the same kind of stuff. Machines lack the complexity, in most respects, to justify our departing from technical explanations of their behaviour. We only need to ascribe mental states and intentionality to systems which are not transparent in their functioning, and where we cannot give an accurate reduction base.

10. Conclusion

In Section 7, I showed why 'technical', neuro-biological explanations cannot replace intentional explanations. Neuro-biological explanations account for the occurrence of bodily movements, period, whereas intentional explanations account for actions. And though actions are nothing over and above neuro-biological events, giving a specific sequence of neuro-biological events cannot explain an action because the same action could be realized by any number of neuro-biological events. In Sections 8 and 9, I argued that reasons are physical causes. But when we give somebody's reasons we do not specify exact physical causes. Rather, we describe the agent's antecedent internal states in terms of their tendency to bring out this kind of behaviour rather than that. Which means, we are really giving a functional, dispositional description of their internal states. But just because we specify the relevant internal events in this way does not mean that we do not mean to denote physical events. When somebody does something for a reason, antecedent internal physical events cause their actions. Accordingly, it is quite mistaken to believe that *either* reasons *or* antecedent physical events cause behaviour. There is no alternative to chose from, since reasons are physical events.

One might describe this view as a special brand of eliminativism.[38] And because it endorses ontological reduction one might also call it a reductionist position. Indeed, I have argued that there are only physical events. There are no mental events over and above that. But this does not mean that there are no mental events. Compare: Lightning is a physical event. It is nothing over and above an electrical discharge between cloud-masses. But we must not say that there is no lightning, period. Similarly, we must not say that there are no mental events, period. And so the mental is not eliminated so much as 'naturalized'. One might call this reduction nonetheless. Not much hinges on it. What matters is that on this view reasons-talk is not 'merely a convenience' for picking out the relevant physical states that explain actions. This would imply that it is a dispensable luxury. The fact is, whatever God could do, *we* could not pick out the relevant states for explaining actions if it were not for reasons-talk, and unless our powers of knowledge approach omniscience one day, we never could. Which means that reasons as causes are here to stay.

Notes

[1] I am grateful to Professor Michael Pendlebury for reading several earlier drafts of this article and offering many helpful comments and suggestions.

[2] In this essay I will regard a state's being about something as being identical to its having a certain content, and thus I will regard a state's having content as a mental property. I will also regard a state having a mental property and its being intentional as identical. 'Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world' (Searle, 1991).

[3] Causal chains can, of course, also 'go wrong' outside the agent – as external deviant causal chains. The example in the literature which comes to mind is D. Dennett's: a man intends to kill another by shooting him; the shot misses but the bang causes a herd of cows to stampede, trampling the victim to death.

[4] Note that this is not necessarily Kenny's own view.

[5] In this section, as well as Sections 4 and 8, my treatment of the topic draws heavily from Heil (1992).

[6] p. 38.

[7] I say 'contingently' because the relational property we pick serves to identify 'in this world, under these circumstances', as Heil puts it (1992: 39).

[8] Ibid.

[9] [redacted]

[10] [redacted]

[11] I say 'mind' without thereby wishing to make a distinction from 'brain'.

[12] Heil is not necessarily stating his own view here.

[13] As Heil argues (24), 'intrinsic' can be equated to 'internal' provided we mean 'logically' or 'conceptually' internal, not 'inside the body'. Thus, my left arm's being longer than my right arm is an extrinsic, relational property of my arm, but it is an intrinsic (dyadic) property of the single complex entity — my body.

[14] Note that Heil is presenting the case for externalism. This is not necessarily his own view.

[15] Note that 'causal connections' need not only mean causal history, but also causal propensities.

[16] I do not want to commit myself to the view that only the causal history of a state determines its content. Its normally bringing about certain behaviour may be just as relevant. But, to keep on track with the main issue I will skirt the issue here.

[17] Actually, I will be advocating a special brand of eliminativism later. I will argue that content is ontologically reducible, though not epistemologically. Thus the preservation of content is not merely pragmatic, it is necessary. But I want to keep to the issue at hand, for now.

[18] I rely on Kim's (1989) account here. I was pleased to discover that Heil (1992) did the same.

[19] Michael Pendlebury made me aware of this.

[20] Heil gives us good reasons to be suspicious of this criticism, and I will raise his considerations in Section 8.

[21] For convenience's sake, I will regard a mind's undergoing an event as being a state of the person.

[22] Professor Mark Leon made me aware of an objection to this counter-example. Tulip and tulip^1 are not really intrinsically identical. Plants are 90% water, and so the former is composed mainly of H20 and the latter of XYZ. Though this objection does not prove that extrinsic properties are causally relevant, it does show that the possibility cannot be excluded that there is a relevant *intrinsic* difference between the two plants in the example. In that case the conclusion would no longer be valid that the presence of sunlight and water for the plant and sunlight and twin-water for plant *are* not causally relevant. Therefore a comparison between the growth of the two plants with *absolutely certain* conclusions is no longer possible. But the objection is not very serious. We have merely raised a practical difficulty: can we come up with an example in which there really are no intrinsic differences when there are extrinsic ones? Not much hinges on this. So long as we can imagine that this were possible, we have all we need to make the important point which follows.

[23] One may add that causally relevant properties are shared, functionally specified internal properties. Cf. my discussion of disjunctive properties in Section 8.

[24] In this section I depend heavily on Lewis.

[25] Naturally, the explanation citing organisms is as subject to possible criticisms as the social worker's pollution explanation. There is always more detail to be filled in, there are always unknown relevant facts.

[26] It is, of course, quite possible to answer this question adequately without mentioning liquidity. But I will take it for granted that it is necessary.

[27] As similar differences matter in all other discourses. In the interpretation of statutes, the rule 'inclusio unius est exclusio alterius' applies. (The express mention of the one is the exclusion of the other.) For example, if there is a law prohibiting 'the importing of drugs, wild animals and exotic fruits', then domestic animals and normal fruits may be imported. The implication lies in the express mention of specific items, which means that we may assume that the class which includes these

items as a sub-group is not meant.

28. Michael Pendlebury made me aware of this.
29. See my discussion of intrinsic properties in Section 3.
30. Note that the same point that was noted (Section 4) regarding the compatibility of externalism and token identity applies in the case of type identity, The reduction base for mental properties includes *both* the physical property of an event (denoted by the disjunctive specification) and its relational properties.
31. It may sound as if I am therefore eschewing an ontological reduction, but this is not so. Rather, the picture that emerges is that mental and physical properties are in the first place only distinguished on the linguistic level. What kinds of properties are there specified is unclear, and the properties designated in both physical and mental language remain mysterious to us. And as Heil has suggested, physical properties are no less mysterious than mental ones. (MBP, p. 19–20).
32. I here concur with Searle, and will take it for granted that a system cannot reason just in virtue of the fact that it manipulates formal symbols.
33. I will take this for granted even though one could argue that it represents kinetic energy in the bimetallic strip, or a variety of other things, including nothing at all. See the relevant discussion in Section 3.
34. The point is, we want to get as close as possible to the idea that the machine *needs* food.
35. My discussion of explanation in Section 6 is relevant here.
36. This conclusion might actually be supported by neurobiologists one day. They might find that whenever there is water in my vicinity, I go into one of n states.
37. This we must do because of the way in which the difference between rational and non-rational creatures has been described. Searle and Dretske's rational creatures 'recruit states in virtue of their semantic properties'. Working within this framework, an explanation must show in virtue of what a state was recruited.
38. As Heil describes his position, which mine closely follows.

Bibliography

Davidson, D. 1970. Mental events', in Foster and Sanson, *Experience and theory*. Amherst, Mass.: University of Massachusetts Press.

Dretske, F. 1990. Does meaning matter?, in *Villanueve, Information, Semantics and Epistemology*. Cambridge, Mass: Basil Blackwell.

Heil, J. 1992. *The nature of true minds*. Cambridge: Cambridge Univ. Pr.

Heil, J. 1993. *Mental causation*. A. Mele (Ed.). Oxford: Oxford Clarendon Press.

Heil, J. *The mind-body problem* (unpublished).

Kim, J. 1989. The myth of non-reductive materialism. *APA Proceedings*, **63**(3), Presidential Addresses.

Lewis, D. 1986. Causal explanation', in *Philosophical Papers*, Vol. 11. Oxford: Oxford Univ. Pr.

Pendlebury, M. 1994. Content and causation in preception. *Philosophy and Phenomenological Research*, **LIV**(4).

Putnam, H. 1975. The meaning of meaning, in K. Gunderson, *Language, mind, and knowledge*. Minneapolis: University of Minnesota Press .

Searle, J. 1991. *Minds, brains and programs*, in Rosenthal, *The nature of mind*. Oxford: Oxford Univ. Pr.