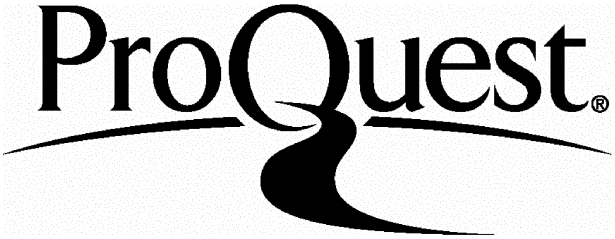
On the Alleged Impossibility of Understanding Consciousness

Submitted by James MacKenzie Garvey

for the degree of PhD in Philosophy

University College London



Abstract

Central to contemporary philosophy of mind are two questions: what is consciousness, and how is consciousness related to the world? This thesis is concerned with sceptical responses to such questions, responses that take a number of forms.

Some claim that empirical data undermine the concept of consciousness, such that it makes sense to say that 'consciousness' will go the way of phlogiston, signatures and spirits. Others argue that consciousness is beyond our ken; our cognitive faculties, in other words, are not up to the job of understanding consciousness. Still others maintain that consciousness is simply not amenable to the kind of understanding we seek.

This thesis is an attempt to clarify these positions, and others like them, and determine whether or not there really are good reasons for thinking that consciousness cannot be understood. In the end, I hope to show that none of the arguments offered by those sceptical about the prospects for understanding consciousness are convincing.

Table of contents

Acknowledgements

Nearly three years into the writing of this thesis the house I live in was burgled, and the thieves made off with, among other things, the only electronic copies of my thesis in existence at the time. This gave me the unusual opportunity to rewrite whole chapters with something like the benefit of hindsight, and I hope the thesis as it stands is better than the one that was lost. I suppose that something of a debt is owed to those responsible for the theft.

I have benefited from the helpful comments, criticism and advice of two supervisors during the writing of this thesis: Professor Ted Honderich, until his retirement, and Dr. Tim Crane thereafter. I am very grateful to Professor Honderich, not only for his time and help in the production of this thesis, but also for giving me some idea of what doing philosophy really is. I also owe a great debt to Dr. Crane, who read and commented on two versions of this thesis and provided many helpful suggestions which issued in a number of substantial improvements. If there is anything true in this thesis, it is due to their efforts. I also owe a substantial debt to my mother, Judith Garvey, and my grandmother, Yolonne MacKenzie, for considerable moral support. And I am indebted to Sophie Davies, who has helped me more than I can say.

This thesis is dedicated to the memory of my father and the memory of my great aunt.

Introduction

Central to contemporary philosophy of mind are two questions: what is consciousness, and how is consciousness related to the rest of the world? This thesis is concerned with what is, from a certain point of view, a particular kind of sceptical response to these questions. The aim is to determine whether the sceptical response is warranted by evaluating the arguments that underwrite various versions of it. Something more needs to be said, if only in a preliminary way, about both the questions and the sceptical response to them.

When a person is awake it is said that she enters into or experiences states of consciousness. She feels the rain and cold wind on her skin, say, wonders why the busses always run late, and imagines owning a car. Such states fill our lives; in a sense they are our lives -- our inner lives of hopes, beliefs, worries, sights, sounds, headaches, jealousies, fears, loves, and on and on. It is admitted in many contemporary books on the philosophy of mind that examples like this one are about the best we can do when it comes to picking out states of consciousness. The states themselves are certainly familiar enough to us, but this sort of inner ostension facilitated by paradigm cases is where many accounts of consciousness begin and end. The debate moves on to other matters, and consciousness itself is left largely uncharacterized.

Sometimes certain expressions, often in conjunction with examples, are used in an effort to say what consciousness is. 'Point of view', 'raw feels', 'subjectivity', 'awareness', 'qualitative states', 'phenomenal states', 'wakefulness', 'attention', 'what it's like' and 'how it is for a subject' are all expressions in this neighbourhood. No doubt these expressions and others like them do something to focus our attention on the problem area, but a case could certainly be made for the view that what we have here are synonyms and approximations, certainly not full-blown, satisfying, philosophical accounts of the nature of consciousness. Something more is probably needed than examples and expressions like these if we are to come to an understanding of consciousness and answer the first question under consideration.

It is possible, though, that we will never be able to say just what consciousness is, that we cannot have a satisfying, philosophical account of consciousness. The difficulty might be conceptual; in other words, we might be precluded from an understanding of consciousness because of certain problems with our concept of consciousness and its relation to the world. It might be the case that our concept of consciousness does not pick out a unified phenomenon amenable to the kind of understanding we are after. Even worse, the concept might not refer at all. In the first case, our position would be analogous to that of a confused person

trying to come to a deep understanding of a more or less arbitrary set of things, like the set of objects within a mile radius of a pencil. States we currently group together as conscious states might share nothing very interesting in common, nothing upon which to build an understanding. In the second case, our position would be analogous to that of a person attempting a philosophical account of unicorns. States we currently call conscious might not really exist at all. In both cases, there just might not be an understanding of consciousness in the sense hoped for to be had in the first place.

There is at least one more sense in which we might not be able to say what consciousness is, and, again, the problem is conceptual. It may be that there really is a genuinely unified phenomenon in the world, and we are, in some sense, pointing to it with our current conception of consciousness, but our concept forming capacities are such that we can never bring the thing into clear view. In other words, the problem might be with our ability to think along the lines necessary for a real understanding of the nature of consciousness. Perhaps evolution did not gear our intellectual faculties for the job. In any case, it could be that our capacity to ask questions about the nature of a thing overreaches our ability to form the concepts required for a satisfying answer. This could be true of the question, what is consciousness.

Similar considerations apply to our efforts to answer the

second question, how is consciousness related to the rest of the world -- the mind-body problem, on the usual reading of it. Certain events and states -- the ones that we point to when we give examples of thoughts, feelings, headaches and the rest or use certain expressions like 'how things are for a subject' -- seem to stand in lawfully correlated relations with events and states of the brain. We know this from results in the neurosciences, results apparently corroborated in our everyday lives when we drink too much beer, suffer a knock on the head, and so on. What exactly is the nature of the relation between conscious events and brain events? How are thoughts and feelings brought about in a physical world?

If it is true that our conception of consciousness fails to pick out a unified phenomenon or fails to refer at all, then, clearly, we are not going to arrive at the kind of answer to this question that many hope to have. If our concept is defective in certain ways, in other words, there is a sense in which we are not asking a proper question at all. Imagine a shaman in the grip of a theory explaining epilepsy in terms of demonic possession. He might wonder how demonic possession issues in the seizures he observes. It is clear that he is not going to arrive at the sort of answer he seeks, because there are no demons to understand. If our conception of consciousness fails to pick out a phenomenon amenable to explanation, then the mind-body problem is on a par with the shaman's demon-seizure problem. No understanding is there to

be had -- not in terms of demons, anyway -- and if the skeptics are correct, not in terms of consciousness either.

Just as in the case with our first question about the nature of consciousness, if our concept forming capacities are not up to the job of solving the mind-body problem, then we cannot arrive at the kind of understanding we hope for. It is possible to believe that 'consciousness' refers and a solution to the mind-body problem exists, but if the concepts needed to satisfy our curiosity are not available to us in virtue of the nature of our cognitive economy, then we cannot solve the mind-body problem. Again, it would be true that consciousness could not be understood.

Some response to sceptical views such as these is certainly warranted. If some ground for resisting the sceptical arguments cannot be found, then a number of disturbing consequences follow. Not only can we expect little or no progress in the philosophy of mind as we know it -- arguably, the philosophy of mind is dead if the skeptics are right -- but problems in other areas of philosophy become far less tractable if it turns out that consciousness cannot be understood. So what form might an adequate response take?

One possibility is advocated by John Searle. He writes,

> How...would one go about refuting the view that consciousness does not exist? Should I pinch its adherents to remind them that they are conscious? Should I pinch myself and report the results in the

Journal of Philosophy? (Searle, 1994, 8)

Though there is much to recommend this response -- it makes a certain point expressed often with respect to sceptical views -- it is clearly lacking in persuasive power. Few are likely to be convinced. A better response, it seems to me, is a careful consideration of the best arguments offered in support of sceptical conclusions.

In the following chapters I plan to consider six arguments from the work of Patricia Churchland, Georges Rey, Kathleen Wilkes, Frank Jackson, and Colin McGinn, all issuing in some version of the conclusion that consciousness cannot be understood. The arguments themselves come from different quarters -- variously invoking empirical results, evolutionary theory, ordinary language, thought experiments, and general observations about our cognitive capacities -- and there is no easy way to classify them or bring them together into one coherent view without doing violence one or more of them. Therefore I will consider each argument more or less on its own and try to determine, in each case, whether or not its conclusion is warranted. In the end, it is hoped, we will be able to come to some general conclusions about the prospects for understanding consciousness.

1. Consciousness and Conceptual Change

Perhaps the best place to begin a consideration of sceptical views of consciousness in the sense just outlined is with arguments for the claim that consciousness cannot be understood because it does not exist. On the face of it, the claim that there is no such thing as consciousness seems incredible, but if we reflect on past efforts to understand something with concepts that fail to refer, the claim is not as farfetched as might be supposed. It may be that evidence from the sciences leads to the conclusion that the concept of consciousness ought to be replaced by an entirely different concept or concepts, just as evidence might have persuaded alchemists that something other than phlogiston is needed if burning is to be understood. In the end, evidence indicated that phlogiston does not exist -- perhaps the same is true of consciousness.

This position is taken up by Churchland, whose dim view of the categories of folk psychology, in particular states of consciousness like the ones that interest us, is well known. (See, for example, Churchland, 1980, 1982, 1983, and 1992.) An early article often cited in her later works, called 'Consciousness: the transmutation of a concept', is perhaps the strongest statement of her views regarding the prospects for understanding certain mental states and events in terms of the concept of consciousness. There, she argues that three aspects of what she characterizes as 'the orthodox concept of

consciousness' are endangered by recent empirical findings. Those aspects are: 'the alleged transparency of the mental, the supposed unity of consciousness and the idea of the self, and the allegedly special relation thought to obtain between language and consciousness'. (Churchland, 1983, 80, also Churchland, 1992)

The method Churchland uses is a little unusual in a philosophical context. Her aim is to show that a certain conception of consciousness is refuted by empirical results in, among other things, the brain and behavioural sciences. This calls for a response that is a little unusual as well. At first, we will have to spend some time considering the data, and we will have to take Churchland's brief remarks concerning the traditional conception at face value. Trying to pin down her understanding of each aspect of the so-called traditional conception of consciousness while attempting to determine how each empirical result works to undermine it is not only too cumbersome an approach, it also robs Churchland of the strongest statement of her position. Until all of the data are before us, and Churchland's conclusion is strongly formulated, it is premature to begin interpreting and evaluating her claims.

So the first section of this chapter involves the statement, in more or less her own words, of each of the three aspects of what she takes to be the orthodox conception of consciousness, followed by a recapitulation of the data

relevant to each one. In the second section, I will consider her conclusion in some detail, focusing in particular on her suggestion that the concept of consciousness will go the way of other dubious concepts like phlogiston. Then will we be in a position to return to the data and interpret the aspects she targets. Finally, in the third section, I hope to show that her conclusion is not a reasonable one. Despite her claims, I will argue, there is no reason to believe that consciousness cannot be understood.

1.1 Consciousness and the Evidence

Transparency

Churchland understands the transparency of the mental as 'the venerable dogma that one's mental life is self-intimating and introspectively available.' (Churchland, 1983, 80) She cites the following nine empirical findings as evidence against the transparency of the mental.

(1) Experiments seem to support the conclusion that stimuli beneath the threshold for conscious awareness have effects on what is normally characterized as behaviour that is under conscious control. In particular, subliminal stimuli have been shown to influence such things as deliberate word choice, problem solving and preference under a variety of conditions. (Shevrin, 1973; Kolers, 1975; Lackner and Garrett, 1972; Zajonc, 1980)

(2) Cognitive psychologists like Chomsky (1965) regularly

postulate a considerable amount of unconscious processing in explanations of intelligent behaviour, such as speech acts. All of this processing, like subliminal influences, escapes introspective notice.

(3) Social psychologists find that some judgments are based on considerations other than those cited by subjects as the basis for their preferences. For example, it has been found that evaluators regularly prefer job candidates they expect to interview over those they do not plan to interview, even though the evaluators do not cite this prospect as a reason for preferring a given candidate. (Nisbett and Wilson, 1977)

(4) It has been found that the blind sometimes use echolocation as the basis for some of their spatial judgements. (Griffin, 1974) Their ability to navigate a room diminishes when their ears are blocked, but they claim beforehand that echolocation does not figure in their ability.

(5) Subjects are more likely to characterize individuals in photographs as friendly, warm and appealing if the pictured individuals have large pupils, even though pupil size is not cited as a basis for their judgments. (Hess, 1975)

(6) Blindsight is a condition resulting from damage to the visual cortex. Blindsighted subjects report blindness in some part of their visual field but do better than chance when asked to guess about the presence or absence of stimuli presented there. (Weiskrantz et al., 1974) 'Blindsight in

humans', she claims, 'puts paid to the idea that perceptual judgements require consciously available perceptions.' (Churchland, 1983, 82)

(7) Some individuals with a condition called 'blindness denial' are totally blinded by damage to the visual cortex but apparently do not know that they cannot see. (Critchley, 1979) She claims that our everyday understanding of seeing and being aware of what one sees leads us to suppose that anyone who does not see cannot help but know about it. However, she concludes that 'blindness denial teaches us that we ought to be prepared to revise our common-sense beliefs about awareness'. (Churchland, 1983, 83)

(8) There are also what Churchland calls 'homey' examples: driving on automatic pilot and daydreaming while simultaneously 'cranking out proofs for freshmen.' (Churchland, 1983, 83) Though she admits that it is not yet clear how we are supposed to think of these and other examples of intelligent activity that escape introspective notice, the data do not indicate 'that the best or most productive model for understanding mental activity *in general* is the one prevailing in folk psychology.' (Churchland, 1983, 83)

(9) Finally, she discusses confabulation, instances in which a subject seems to invent and then believe pseudo-reasons for some of the choices and judgments he makes. An example from the literature (Gazzaniga, 1978) is used to make the point. A split-brain patient, a person whose corpus

callosum is severed, receives the following visual stimuli: a picture of a snowy scene is flashed to the right half of his brain and a chicken's claw to the left. When the patient is asked to select cards appropriate to the image from an array before him, the right hand chooses a chicken's head (apparently, to go with the chicken's claw) and the left chooses a shovel (apparently, to go with the snowy scene). When asked to explain, the hemisphere in charge of speech, the left one, confabulates: '"I saw the claw and I picked the chicken, and then the shovel because you have to clean out the chicken shed with a shovel"....'(Churchland, 1983, 84) According to Churchland, the left hemisphere of a split-brain patient has no access to the visual stimuli received by the right (in this case, a snowy scene), but the left side nevertheless explains the right's choice of the shovel by confabulating a reason: something suitable about a chicken shed that seems to mesh nicely with the chicken claw. There is, she claims, no apparent puzzlement on the patient's part about what the left hand was doing.

She reports that ordinary individuals exhibit confabulatory tendencies as well. When presented with roughly identical pairs of tights and asked to compare them qualitatively, subjects typically prefer the pair on the right, though they explain their choices in terms of nonexistent differences in strength, sheerness, colour and the like. (Nisbett and Wilson, 1970) There is no evidence, she

claims, that 'confabulation is pathological; rather it appears to be a normal part of theorizing about persons and their behaviour, oneself and one's own behaviour.' (Churchland, 1983, 85)

In light of all of these data, Churchland concludes, the alleged transparency of the mental,

> ...seems at last to be ready for consignment to the museum of quaint and antiquated myths about how humans work.... [A] great deal of intelligent and sentient activity, perhaps the lion's share, goes on without benefit of self-conscious awareness, and that so far from this being a pathological condition, it is in the nature of the case. (Churchland, 1983, 80-1)

We will return to this aspect shortly. For now, let us continue with a consideration of the set of data Churchland believes undermines the second aspect of the orthodox conception of consciousness.

The unity of consciousness and the idea of the self

For now, we might understand this aspect as a commitment to the existence of a single, seamless something -- a conscious unity or the awareness of a self, perhaps -- in control of deliberate behaviour. More will be said about this

in the next section. Given the way Churchland's consideration of introspection dovetails into her discussion of the self and the unity of consciousness, and given the fact that her treatment of the latter is comparatively brief, it is reasonable to suspect that, by Churchland's lights, evidence adduced and remarks made with regard to the former topic suggest something about the latter.

Her discussion of transparency includes evidence against the claim that we have direct cognitive access to the underlying mechanisms which control some behaviour, and 'if our conception of control changes, then our conception of the self changes *pari passu*'. (Churchland, 1983, 85) Thus, for her, reflections on the possibility that some of our seemingly deliberate actions actually have unconscious precursors suggests something about the nature of the mechanism or mechanisms controlling those actions. If at least some of the actions that one takes to be deliberate are actually the result of unconscious processing, then perhaps thinking that there is a self or a unified consciousness behind behaviour is illusory too. Churchland writes,

> ...if unconscious states can figure in choice and judgment, then the old dogma is beleaguered also in its insistence that consciousness and control go hand in glove. Control may turn out to be only incidentally and occasionally connected with

awareness, and the dynamics of human behaviour may turn out to bear but passing resemblance to the prevailing notion of the conscious self in control of cognitive activity and deliberate behaviour. (Churchland, 1983, 84)

She cites three sets of data.

(10) The study of split-brain patients suggests that the control of different capacities is handled separately by the two hemispheres. Churchland does not report the details of these studies, but she does observe that some theorists (Gazzaniga and LeDoux, 1978) are led by their discoveries to conclude that 'the mind is not a unified psychological entity, but a "sociological" entity, in the sense that *control is distributed*...[and] the conception of the unified self is at bottom illusory.' (Churchland, 1983, 84)

(11) She cites studies of dissociation phenomena, such as fugues and split-personalities. (Hilgard, 1977) A fugue is an episode in which a person appears to exhibit characteristic responses to stimuli, but, after the episode, recalls nothing. It is not exactly clear why fugues, which are taken by many as simple amnesiac episodes, should count as evidence against a unified self. Nevertheless, she understands both fugues and split-personalities as instances of control fragmenting in various ways.

(12) Churchland claims that it is not just

psychopathological phenomena that are telling. She maintains that even the commonplace of doing several things at once -- to use her example, listening to a conversation at a party while formulating something to say, keeping a plate of food and a glass of drink steady, and remaining all the while on the lookout for the host -- 'fails to square with the idea of unity of control'. (Churchland, 1983, 84)

Echoing some of her concluding remarks concerning introspection, she admits that it is not clear how such cases should be characterized. She writes,

> Must there be a supervenient controller who divides subsidiary control, or is it possible that a strict hierarchial model of control is wrong, that control shifts and moves in a quite different dynamic?...[W]e should be ready to find that a strictly hierarchical model of control may be as inappropriate to the brain's organization as the model of intelligent creator was to the organization of Nature. (Churchland, 1983, 84)

Again, we will return to this aspect and her conclusion concerning it in a moment. Let us first consider the third and final aspect of the orthodox conception that, Churchland alleges, is endangered by empirical findings.

The special relation between thought and language

The third aspect of the orthodox conception of consciousness under consideration is the special relation thought to exist between thought and language. Churchland claims that this is probably the most opaque of the three aspects of the concept of consciousness, primarily because it is so difficult to find a substantive expression of it. She nevertheless tries to press the thesis into focus by citing claims advanced by several different thinkers. Gazzaniga and LeDoux (1978) claim that language somehow creates consciousness, or at least makes human consciousness what it is. Eccles (1977) holds that the interface between the nonphysical mind and the brain is the part of the brain associated with the use of language. And Popper (1977) understands consciousness as an emergent property, with language identified as the vehicle of emergence. What all of this amounts to is never exactly made clear -- no doubt a reflection of Churchland's difficulty in making sense of this aspect of the concept. However, she claims that the idea that language is essential for consciousness, which all of these views express in one way or another, 'is widespread, and may perhaps be granted the status of conventional wisdom'. (Churchland, 1983, 85)

The notion is put in doubt, though, by the following:

(13) Ethologists maintain that higher animals are conscious despite being unable to speak, and similar claims

are made about preverbal children and nonverbal adults.

Churchland identifies another, slightly modified version of the orthodoxy's claim that language is essential to consciousness. According to this formulation, animals and other nonverbal creatures are conscious, but there is a distinction to be drawn between the sense in which they are conscious and full-blown, human self-consciousness. Language, on this modified view, is regarded as essential for certain representations, and these representations are, in turn, essential for self-consciousness. So though nonverbal creatures are conscious, they are not self-conscious in the way that language users are. However, Churchland cites the following as evidence against this modification.

(14) '[T]he complexity of behaviour in many species of animals invites, nay *requires*, the postulation of an internal system of representations.' (Churchland, 1983, 88) Further, given our evolutionary ties to nonverbal animals and the number of structural and developmental parallels that exist between us and them, we have reason to think that they are as conscious as we are, and this includes crediting them with a conception of self akin to our own. She cites tests involving chimpanzees exhibiting self-recognition behaviour in mirrors as suggestive of the possibility that they too have self-consciousness. (Gallup, 1977)

She concludes, once again, that it is not yet clear how we should think about the various ways in which brains manage

to represent the world, other brains, and themselves. However,

> What we should avoid in trying to solve the mystery is the idea that linguistic representation is the prototype for representation generally. Accordingly, we should seek models of representation which operate on radically different principles....My point is that for all we know, the differences between my conscious states and those of á gorilla may be no more differences in *kind*, and no more significant, than those obtaining between a medieval serf and a modern man steeped in biological and physical theory.... (Churchland, 1983, 92)

Churchland's conclusion

So, by Churchland's lights, a certain conception of introspective access, a particular view of the self, and a special relation thought to hold between thought and language are, in some sense, in conflict with data. This leads her to draw a general conclusion about the future of what she calls 'the orthodox concept of consciousness'. She writes:

> It sometimes happens in the history of science that well-used, highly entrenched, revered and respected concepts come unstuck. That is, under the suasion

> of a variety of empirical-cum-theoretical forces, certain concepts lose their integrity and fall apart. Their niche in the theoretical and explanatory scheme of things is reconstructed and reconstrued, and new concepts with quite different dimensions and dynamics come to occupy the newly carved niche. (Churchland, 1983, 80)

Churchland's claim is that, given the empirical evidence adduced, 'a similar fate may befall concepts respected and revered in our own prevailing conception of how humans work and the concept on which I mean to focus is consciousness.' (Churchland, 1983, 80)

1.2 The Orthodox Conception and its Alleged Fate

All of this raises a number of questions of interpretation which must be dealt with before we can say whether or not Churchland's conclusion is reasonable. We know that she claims that a certain conception of consciousness is 'falling apart', 'coming unstuck', and 'losing its integrity'. There are two questions here alone. First, is there a clear, non-metaphorical way to express this conclusion? We will take up that question first. Another question that threatened to surface throughout the first section is this: what is the nature of the conception that is allegedly 'falling apart', 'coming unstuck', or otherwise in jeopardy? She claims that

it is the orthodox conception, that the components she lists have the status of conventional wisdom. Is this really the case? At the end of this section, it is hoped that both of these questions -- how are we to understand her conclusion, and what conception of consciousness is operative in her thinking -- will have satisfactory answers.

Conceptual crumbling

What exactly is it to say that the concept of consciousness is falling apart or crumbling? Her claim is that the data reported put a particular kind of pressure on the orthodox concept of consciousness, pressure that results in erosion, a loss of integrity, crumbling. What is crumbling away, presumably, are the three aspects of the concept under consideration. For the claim that the concept itself is falling apart, these aspects must be crucial parts of it -- the concept stands or falls with them. Insofar as they fall away, the concept itself is weakened, and this underwrites an inference to the conclusion that the concept will be replaced by different ones that slice up what we call the mental in a different way.

It would be useful to have a way of thinking about the possible fate of the concept of consciousness she has in mind that does not depend on the 'crumbling' metaphor used. The following passage is instructive:

> The 'spirits' and 'principle' of alchemy, the 'crystal spheres' of pre-Galilean astronomy, 'demonic possession' of Medieval medicine, 'phlogiston', 'ether', and 'signatures', are now nought but dry bones of an earlier intellectual ecology. The theme of this paper is that a similar fate may befall concepts respected and revered in our own prevailing conception of how humans work, and the concept on which I mean to focus is consciousness. (Churchland, 1983, 80)

Clearly, her claim is that what happened to the dubious concepts listed might be happening to the orthodox conception of consciousness as well. Because invoking suspect notions from the past is a common ploy among those sceptical about the prospects for understanding consciousness and because it will help us to understand Churchland's conclusion in a non-metaphorical way, let us linger over the fate of one of the concepts she lists.

Consider phlogiston. Before Lavoisier, the behaviour of burning objects was partly explained by the loss of a diaphanous substance, phlogiston. The notion was undermined when it was discovered that some substances actually gain weight when burned. The problem was not simply that the theory in which phlogiston figured failed to predict this result. Worse, the result should have been impossible. The

existence of an object that gains weight after combustion contradicts something fundamental to the concept of phlogiston. Whatever else phlogiston was thought to be, it was ultimately something that all burning things give off, and this, for phlogiston theorists, required a loss of weight after burning. For some time phlogiston theorists salvaged what they could by claiming that phlogiston has negative weight, but eventually burning came to be understood in an entirely different way precisely because some things gain weight when burned. The new conceptions that took the place of phlogiston in theories of burning, oxidation and reduction, not only made sense of data anomalous for phlogiston theorists, but managed to explain other phenomena too, like rusting.

This example might be used to put some flesh on Churchland's claims about the relationship between recalcitrant data and conceptual change. Presumably, the discovery of weight gain after combustion put in motion the empirical and theoretical forces that Churchland claims sometimes undermine theoretical concepts. A fundamental part of the concept of phlogiston was contradicted by experimental results. The simple fact is that there was an empirical disproof of its existence.

The replacement of phlogiston theory by the theory employing oxidation and reduction is an instance of, to use Churchland's language, new concepts with different dimensions

occupying a newly carved niche in our understanding of the world. A very good case could be made for the claim that at least most of the other concepts Churchland lists were undermined in just the same way as phlogiston: some empirical disproof led to the conclusion that the thing in question just did not exist, and conceptions of other things moved into the explanatory structure. No doubt Churchland would agree that conceptual change of this kind is a very complex matter, but, arguably, the prime mover underlying this sort of change is empirical disproof. If it can be shown that consciousness now faces recalcitrant data issuing in the disproof of its existence, much like phlogiston did, then the conclusion that consciousness will go the way of phlogiston looms.

At least part of what Churchland is doing, then, is setting up an analogy between the concept of consciousness and dubious concepts like phlogiston. Let us stick with phlogiston for the moment and follow the analogy through. It seems to have three principal components. First, just as phlogiston had a clear place in a theory of combustion, with well-defined aspects that might come into conflict with data, so consciousness has a clear place in our understanding of how humans work, with well-defined components that might come into conflict with data. Second, the concept of consciousness, like phlogiston, is encountering or has encountered empirical disproof; therefore, not only is its place in our understanding of ourselves threatened, in general there is

reason to believe that the concept just does not refer. Third, if this much is true of the concept of consciousness, given what happened to phlogiston, we arrive at the conclusion that consciousness, like phlogiston before it, will be replaced by conceptions with different dimensions, conceptions of different things altogether. If consciousness really is like phlogiston and the data undermine it in the same way, then Churchland's conclusion here is clear: there is no such thing as consciousness.

Although the bulk of her paper is concerned with the second and perhaps most crucial part of the analogy -- the claim that the concept of consciousness faces empirical disproof -- let us consider each part in turn, with a view towards determining whether or not the concept of consciousness really is facing the same fate as did phlogiston.

The ordinary conception of consciousness or a modification of it?

First, if the concept of consciousness is a theoretical notion like the concept of phlogiston, then it has certain fundamental aspects which, if undermined, weaken the integrity of the concept and threaten its explanatory role. Hence Churchland's focus on the three aspects of the orthodox conception. Is the concept of consciousness analogous to the concept of phlogiston in this respect?

How this question is answered depends largely on which conception of consciousness Churchland is attacking, and there are many possibilities. There is the ordinary conception at work in everyday predictions, descriptions and explanations of some mental states and actions. There are also many modifications of this conception operative in philosophy and the behavioural and brain sciences. The fact that it makes sense to ask which concept Churchland is attacking indicates that there is certain a weakness in the first part of the analogy. Even though Churchland calls the concept she has in mind 'the orthodox conception', the meaning of the term 'consciousness' is not, to take the expression literally, currently fixed an orthodoxy, a set of practitioners allegedly in possession of special knowledge, as other terms like 'phlogiston' and 'ether' no doubt were. There is no standard practice like those she cites, alchemy and medieval medicine, in a position to delineate the term.

Further, it might even be claimed that the aspects of the concept that she does attack are not 'highly entrenched, revered and respected', (Churchland, 1983, 80) with the 'status of conventional wisdom', (Churchland, 1983, 85) but objects of contention across largely divided fields of inquiry. Much depends upon how the aspects are interpreted, but the fact that the aspects require interpretation at all suggests that they are not highly entrenched, revered and respected. If this is true, then the situation is largely

disanalogous to the case of phlogiston theory and similar, dubious theoretical conceptions. Attacking *the* concept of consciousness by identifying anomalous data does not make much sense in such a context. This is not to claim that there is *no* useful conception of consciousness, either to be had or to make do with at present. The point is just that there may not now be a clear, orthodox conception of consciousness, as Churchland suggests.

So which conception of consciousness is Churchland targeting? We know from her other writings that she regards the ordinary conception as dubious, so it seems tempting to interpret her as focusing on it. The ordinary conception, as she understands it, figures into folk psychology, 'a rough-hewn set of concepts, generalizations, and rules of thumb we all standardly use in explaining and predicting human behaviour.' (Churchland, 1992, 229) If someone comes home from a brisk bike ride, for example, we explain and predict his ensuing drinking and eating behaviour in terms of his awareness of inner states like hunger and thirst, and we credit him with certain beliefs about the ways in which such inner states are satisfied. Explanations and predictions like this one presumably make use of an ordinary conception of consciousness that Churchland might have in mind.

There is considerable textual support for interpreting her as attacking the ordinary conception of consciousness that figures into such generalizations in this article as well.

For example, she says that the evidence shows that 'we ought to be prepared to revise our common sense beliefs about awareness.' (Churchland, 1983, 83)

There are a number of problems with this interpretation, however. If Churchland's point is just that the ordinary conception will not do for philosophical and scientific purposes, then her argument is not very interesting. Of course the ordinary conception is not good enough for serious philosophical and scientific work. Of course it has difficulty accommodating phenomena like blindsight and blindness denial. Who would have thought that it would not? The ordinary conception is good enough for ordinary purposes, for predicting, explaining and describing mental events and actions to a level of precision that is adequate for getting on with the business of everyday life. Evidence for that is ubiquitous. Ordinary life, though, is not a carefully designed experimental situation, nor does it require a concept that takes account of split-brains and the like.

Churchland does cite some 'homey cases', such as driving home on automatic pilot, about which there is supposed to be some difficulty for the everyday conception. Two points are worth emphasizing here. First, in the vast majority of cases, the ordinary conception does what it does extremely well. Second, the ordinary conception is, as Churchland says, 'rough-hewn'. The fact that it is not precise enough to adjudicate every case is therefore hardly surprising.

Further, emphasizing the fact that the ordinary conception will not do in difficult cases does not necessitate the conclusion that a better conception cannot be had, a conception that is not a radical departure from the ordinary one. The ordinary conception handles many cases well and others less so. If the evidence amassed tells us anything about the ordinary conception, it is that the ordinary conception is inadequate for serious philosophical and scientific purposes. We knew that already. Much more is needed to get from these simple facts to the conclusion that consciousness will go the way of phlogiston.

There is another worry worth mentioning with this interpretation. If Churchland were attacking a conception of consciousness that figures into folk psychology, then it would be a concept that 'we all standardly use in explaining and predicting human behaviour'. (Churchland, 1992, 299) As such, we would expect the aspects of the concept Churchland identifies to be truisms, standardly used by everybody in explanations and predictions of some human behaviour. However, what she describes looks much more specialized, perhaps modifications of the ordinary conception, often characterized in the words of experts in certain fields. It would be a little outrageous to suppose, as we must on the interpretation that Churchland is attacking an ordinary conception we all use, that we all follow Popper and understand consciousness as 'an emergent property [and we

identify language as] the vehicle of its emergence'. (Churchland, 1983, 85)

So perhaps it is best to interpret Churchland as attacking some modification of the ordinary conception, possibly a highly entrenched conception that figures into many accounts of consciousness in the philosophical and scientific literature. If that is the focus of her attack, then at least her conclusion is not a trivial one. Further, this way of reading her strengthens the first part of the analogy with phlogiston. A specialized conception is, almost by definition, not rough hewn, and, therefore, it probably has clear aspects, set forth by experts, that might be undermined by data. Like phlogiston, such a modification also figures into a theory, with a well-defined explanatory role and clear targets that data might undermine.

There is another reason to interpret Churchland as attacking a modification of the ordinary conception. Whether or not the ordinary conception is actually a theoretical construction like phlogiston is the subject of considerable controversy. Churchland might be able to avoid all of this by claiming that her target is not the ordinary conception, but a modification of it that really does figure into philosophical and scientific theorizing about the mind. In addition to side-stepping that debate, this reading of her also makes sense of the fact that she regularly cites experts when characterizing the three aspects of the concept.

It should be noted that reading Churchland in this way does not automatically eliminate worries for the ordinary conception. It might give us considerable cause for concern about the ordinary conception if it turns out that a commonly employed modification of it is in deep trouble, and it seems likely that this is precisely Churchland's point. Perhaps the reason the modification is in danger can be traced to the original conception. That unspoken inference might, in turn, make sense of her sceptical claims about the ordinary conception peppered throughout the article.

Her view, then, seems to be that the modification currently employed by a number of people across a variety of disciplines is facing empirical disproof, and the problem is that our theorizing started with assumptions rooted in the ordinary conception. We need to be prepared to revise our common-sense beliefs about consciousness, she says over and over again, because those beliefs are what led us to the modification in the first place. Whether or not this conclusion about the ordinary conception is warranted will concern us in the final section of this chapter.

So if it is a modification of the ordinary conception under attack, how are we to understand its three aspects? For the first part of the analogy to hold and for her argument to work, the aspects must be crucial to the concept, like three-sidedness is to triangularity and, perhaps, substance-responsible-for-weight-loss-in-burned-objects is to

phlogiston. We shall see whether or not the aspects under consideration really are fundamental to conceptions of consciousness at work in philosophy and the brain and behavioural sciences, and we shall also see whether or not the three aspects are crucial to any conception of consciousness properly so called. If they are, then it will turn out that not just the modification under consideration is in trouble, but every conception of consciousness is in danger of going the way of phlogiston. This is the worrying possibility, for our purposes.

Because Churchland spends more time discussing experimental data than explaining the aspects of the modification, and since the first section of this chapter has given us a rough idea of her understanding of the data, it might be best to consider the aspects in conjunction with the data used to undermine them. How the data are used should give us some idea of the meaning of the aspects in question. We move on, then, to the second part of the analogy, the claim that the conception of consciousness under consideration, like phlogiston before it, has encountered data which threatens its place in the explanatory scheme of things.

First aspect of the modification: transparency

The first aspect of the modified conception is, as we have seen, 'the venerable dogma that one's mental life is self-intimating and introspectively available'. (Churchland,

1983, 80) She also calls it the 'transparency of the mental'. (Churchland, 1983, 80) The words 'mental life', 'self-intimating', 'introspectively available', and 'transparency' are all open to a variety of interpretations. What does Churchland have in mind?

Here is her account of transparency, in full:

> The venerable dogma that one's mental life is self-intimating and introspectively available seems at last to be ready for consignment to the museum of quaint and antiquated myths about how humans work. To be sure, the psychoanalytic literature had worked some wrinkles into the dogma, but on its own, psychoanalytic theory was perhaps not enough to raise serious and systematic problems. After all, one could always say that it was really only the naughty and nasty bits of one's mental life that sometimes escaped introspective notice, not the exalted and more serious business of cognition and sentience. Moreover, it was generally thought that given the right conditions, even the seamier side of one's mental life could be marched out of the occluding shadows. In other words, it was seen as an aberration from the typical transparency of mental life that certain psychological states and processes escaped even the most discerning

> introspective eye. Now, however, there is evidence that a great deal of intelligent and sentient activity, perhaps the lion's share, goes on without benefit of self-conscious awareness, and that so far from this being a pathological condition, it is in the nature of the case. Now for the evidence. (Churchland, 1983, 80-1)

The difficulty here is, of course, that it is not clear which doctrine of introspection she is taking as her target, nor is it obvious what she means by 'mental life'. The major presupposition of this passage is that there is one more or less dominant conception of the transparency of the mental, venerable dogma that need not be rehearsed here. Is that true, and, if so, what is it?

Speaking very generally, so as to beg as few questions as possible, introspection might be taken as the way in which subjects ascertain the nature of, as Churchland puts it, their mental lives. How are we to understand 'mental lives'? Churchland says that some wrinkles were worked into the dogma of transparency by psychoanalytic literature. The problems only seemed superficial, she suggests, because 'one could always say that it was really only the naughty and nasty bits of one's mental life that sometimes escaped introspective notice, not the exalted and more serious business of cognition and sentience.' (Churchland, 1983, 81) The idea seems to be

that before Freud's claim that we sometimes repress the real motivations for our actions, it was generally believed that introspection provides sure access to the whole of mental life. 'Mental life' is being construed very broadly indeed, including within it 'psychological states and processes', (Churchland, 1983, 81) 'intelligent and sentient activity', (Churchland, 1983, 81) and the 'business of cognition and sentience' (Churchland, 1983, 81) credited not just to the conscious mind, but, if the remark about psychoanalytic literature is taken seriously, the un- or sub- or nonconscious mind as well.

Again, begging as few questions as possible, one might identify three general views about the scope of introspective access, the extent to which the mental is transparent, as Churchland is putting it: we have this kind of access to all, some or none of our mental lives. Which is she crediting to the dogmatic view? Calling the doctrine 'the transparency of the mental' rules out the claim that we have this kind of access to no part of our mental lives. The reference to Freudian doctrine suggests very strongly that it is the global view she is targeting, the claim that via introspection one can ascertain the nature of the whole of one's mental life, broadly construed.

Further, much of the evidence that Churchland cites (1-6 and 8) does seem geared to undermine this sort of thinking. Subliminal stimuli cause certain mental events to which we

have no introspective access. Language use seems to require a considerable amount of processing that is phenomenologically blank to us. Pupil size, echolocation and facts about prospective employees might influence our judgments without our knowing about it. Blindsight, on one interpretation at least (and there are several; we will consider others in the final chapter of this thesis), is much like subliminal perception: unconscious visual stimuli have certain effects on perceptual judgments that are unknown to the blindsighted. We also manage to do things like drive home on automatic pilot; which seems very much like sentient activity going on without the benefit of introspective awareness. What this sort of thing is meant to suggest is, one supposes, that even sophisticated mental activity can escape introspective notice, not just the workings of something like edge detectors and other processors contributing to the early stages of our conscious experience, as might be the case in blindsight. In each of these cases, the evidence seems to show that there are mental events or capacities or states that escape introspective notice. Given what this evidence appears to be evidence against, it seems clear that the first aspect of the orthodox concept of consciousness is this: introspection grants access to the whole of our mental lives.

The following question immediately arises: is this claim properly characterized as 'traditional', 'orthodox' and 'dogmatic'? Let us begin with her remarks concerning

psychoanalytic literature and the practice of psychoanalysis generally. If Churchland is right, Freudian doctrine forced wrinkles in a venerable, dogmatic conception of introspection, worries that have become more and more systematic as the evidence against it piles up. This is just not true.

First of all, long before Freud, philosophers and psychologists noticed that at least some mental properties are unavailable to introspection. Leibniz, Schelling, and Nietzsche all anticipate the notion of unconscious mental events, unavailable to introspection, to some extent (Ellenberger, 1970). It seems likely that Pascal's observation -- *le coeur a ses raisons que la raison ne connaît point* -- was available to many thinkers, before Freud and the data Churchland cites. Second, it seems clear that introspection has a long and convoluted history, and even the most cursory consideration of it reveals how difficult it is to find even a proponent of global transparency, much less support for the claim that the global transparency thesis is a dogmatic one.

It is not clear where discussion of introspection originates, though there is reason to locate the beginning as far back as Aristotle's claim that 'the mind too is then able to think *itself*.' *(De Anima, 3.4. 429b 8-9)* The discussion is difficult to follow, but what we find is Aristotle making a case for the claim that the mind can think about itself, not that the mind knows itself in its entirety.

Augustine's *de Trinitate* (10.13, 85) contains a well known passage that foreshadows the Cartesian view that it is possible to focus on the mind, excluding everything that enters it via the senses, and come to a kind of certainty about mental activity. We can, he writes, 'concentrate our attention upon the points which all minds know with certainty about themselves....[that one] lives, and remembers, understands, wills, thinks, knows, and judges.' If anything, there are intertwined claims here concerning the immediacy of and certainty one might have about the kinds of activities the mind engages in, but nothing concerning scope of access. Augustine is saying that we might go wrong in our beliefs about things external to us, but there is something about conscious states that cannot be doubted. There is a special kind of certainty that attaches to the belief that we have certain capacities -- do certain things like think, understand, will and so on -- that does not attach to beliefs about things outside the mind. The existence of conscious states is immediate to us, in a way that the existence of rocks and trees is not. This brings with it, for Augustine, a greater kind of certainty about the fact that one lives, remembers, understands and the like compared to our beliefs about things in the world. Augustine's claim -- that one can have doubts about beliefs concerning objects but know with certainty that one engages in certain mental operations -- is no commitment to the global transparency thesis.

Though Augustine foreshadows Descartes, there is certainly not a straight line from Augustinian thinking to Cartesian thinking. Aquinas, for example, agrees with Augustine that 'the mind perceives itself', but he expressly rejects the claim that introspection is a source of certain knowledge. According to Lyons, for Aquinas, 'introspection was, ordinarily, a mere concomitant mental perception of the basic exercise of mind...peripheral in importance and also in a more literal sense. It was, so to speak, "seeing" out of the corner of the mind's eye rather than with it.' (Lyons, 1986, 2) Though there is agreement that the mind can somehow perceive or reflect on itself, there is, as early as Aquinas, nothing dogmatic about how this claim is to be understood or about the scope of inner awareness.

With Descartes, there is a return to something of the Augustinian view. Descartes resolved to 'embrace in my judgment only what presented itself to my mind so clearly and distinctly that I had no occasion to doubt it.' (Descartes, 1954, 20) Knowledge of everything but the bare existence of states of consciousness -- thinking, believing, doubting, affirming, willing, seeming to see, and so on -- was relegated to a kind of secondary status, ultimately epistemically dependent upon the first certainty of the *cogito*.

It is often said that, for Descartes, all mental events are conscious events and, as such, accessible via the

introspective eye. Here, perhaps, is grist for Churchland's mill, but care is needed. Is it really the case that he held something akin to the global transparency thesis? The class of events in Descartes' conception of mental life is certainly smaller than Churchland's, and this difference is probably reason enough to make one sceptical about attributing belief in global transparency to Descartes. Perhaps it is true that, for him, all mental life is accessible, but he certainly understands 'all mental life' more narrowly than Churchland does. The global transparency thesis, as Churchland understands it, might be certain Cartesian conclusions about the epistemic status of beliefs about our inner lives grafted on to a relatively new conception of mind which encompasses events like subliminal perception, the Freudian unconscious, and so on. If the transparency thesis is just this, then no doubt it is Cartesian in flavour, but it is probably heavy handed to attribute it to Descartes himself.

Variations of Descartes' original view were taken up by many who came after him. Hobbes, for example, made the fruits of inner examination a basis for much political and social reflection: 'Whosoever looketh into himself, and considereth what he doth, and when he does *think, opine, reason, hope, feare* &c, and upon what grounds; he shall thereby read and know, what are the thoughts, and Passions of all other men, upon the like occasions.' (Hobbes, 1968) Notice that this view commits Hobbes to the claim that people have, in general,

like access to the states into which they enter at certain times, but not much follows about the scope or even the certainty of the access.

Brentano held that one could not attempt to learn anything valuable about mental phenomena via the kind of inner observation exemplified in the *Meditations*. Directly observing mental events changes the normal flow of feelings, desires and the like. The task of empirical psychology, he maintained, was to try to perceive mental phenomena passively, out of the corner of one's mental eye, as Aquinas might put it. Wundt attempted to formalize this process by carefully regulating his subjects' inner responses with rigorously controlled experimental situations and elaborate training. James took up the direct, active method of inner observation rejected by Brentano, claiming that '*Introspective Observation is what we have to rely on first and foremost and always*. The word introspection need hardly be defined -- it means, of course, the looking into our own minds and reporting what we there discover. *Every one agrees that we there discover states of consciousness*.' (James, 1950)

With this distinction between passive and active methods of introspecting, the attempt to identify anything like the venerable dogma Churchland has in mind in the history of philosophy is complicated further. It is also worth pointing out that there is nothing in the mature views of Brentano, Wundt or James that suggests a commitment to the global

transparency thesis. If a trend emerges, it is the innocuous claim that subjects can, sometimes by extraordinarily elaborate means, ascertain the nature of some of their mental states. For the experimental introspectionists especially, the route to reliable information about inner states was tortuous, involving a complicated training regime and meticulously controlled experimental situations. This suggests that global transparency was certainly not one of their presuppositions.

Perhaps as a result of the debates amongst scientific psychologists about, among other things, the rules that ought to govern introspective investigation and the conflicting interpretation of its results, introspection at the beginning of the twentieth century found itself in disrepute, and the early behaviourist position arose. According to Watson, 'psychology as the behaviourist views it is a purely objective experimental branch of natural science. Its theoretical goal is the prediction and control of behaviour. Introspection forms no essential part of its methods....' (Watson, 1913, 158) When behaviourists did turn their attention to introspection, it was often treated as a species of, or at least akin to, speech behaviour, that is, a kind of silent speech. Again, nothing approaching global transparency figures into their account of introspection.

The interests of the materialists were largely elsewhere too, and when they treated introspection, it was often only to

undermine the claim that introspective reports were evidence of Cartesian souls. For materialists like Place and Smart, the tactic was to give an account of reports such as 'I have a blurry, red after-image' that was amenable to materialism.

Armstrong gives a fuller account of introspection in terms of a brain scanner. Perception and introspection are just processes that scan different regions: the environment and the brain respectively. Further, both can be in error: 'We can very easily conceive that, in a future where far more is known than at present about the workings of the brain, it would be possible to be quite sure that certain introspections were illusory. I might appear to myself to be angry, but *know* myself to be afraid'. (Armstrong, 1968, 328) If Armstrong is a good example of the materialist position, there is not much ground for attributing global transparency to materialist conceptions of introspection.

No doubt there are other proponents of different doctrines of introspection that have been omitted in this brief account, and clearly more could be said about each of the views considered. The point is that if the global transparency thesis really is a dogmatic presupposition of the orthodox conception of consciousness, one would have expected to find numerous expressions of it, and we have not. It is not clear where else we should look, if not Aristotle, Augustine, Aquinas, Descartes, Hobbes, Brentano, Wundt, James, and proponents of modern materialism. It seems likely that,

far from venerable dogma, the global transparency thesis is a straw man, an easy target not seriously or clearly entertained by any of the philosophers we have considered. If this is true, the relevance of Churchland's treatment of the first aspect of the modification is in doubt. We will return to this worry later. For now, let us consider the second aspect of the modification.

Second aspect of the modification: conscious unity and the self

The second aspect of the conception under consideration is 'the supposed unity of consciousness and the idea of the self.' (Churchland, 1983, 80) The particular doctrine of unity or self that Churchland has in mind emerges in conjunction with the data she considers. It is best to have some text before us if we are to come to any conclusions about what exactly Churchland is arguing against.

As suggested earlier, her treatment of the second aspect is comparatively brief because Churchland seems to be arguing from remarks made about introspection to conclusions here as well. She writes:

> ...if unconscious states can figure in choice and judgement, then the old dogma is beleaguered also in its insistence that consciousness and control go hand in glove. Control may turn out to be only

> incidentally and occasionally connected with awareness, and the dynamics of human behaviour may turn out to bear but passing resemblance to the prevailing notion of the conscious self in control of cognitive activity and deliberate behaviour. (Churchland, 1983, 84)

Although she does not offer a straightforward account of the conception of 'self' or 'unity' at issue, we do have the claim that the dogma under attack says the self is in control of cognitive activity and deliberate behaviour. No new evidence is cited against this claim, but we are presented with the speculations of a few experimental psychologists. 'Whether they are right is an empirical question', (Churchland, 1983, 84) she says, and this is where Churchland's treatment of the self ends. In so far as her remarks here depend upon the dubious conception of global transparency considered in the previous section, her claims may be treated with the same reservations. In other words, if the idea is that a self, however characterized, has global control of all behaviour, even behaviour that issues from the effects of something like subliminal stimulation, then it seems likely from the outset that her claims about the self deserve the same level of suspicion as do her claims about global transparency.

The new data cited in this section -- reports of split-brain patients, split personalities, and ordinary individuals

doing several things at once -- seem to be arrayed in support of a claim not about the self but about the unity of consciousness. The inference here is telling: 'the unity of consciousness is an illusion' (Churchland, 1983, 84) because the data fail 'to square with the idea of unity of control.' (Churchland, 1983, 84)

So far, we have no substantive claim about the nature of unity or the self, but a number of conclusions about control. This is, I think, the crux of things for Churchland. She is not arguing directly against some notion of the self, nor are her claims aimed at the unity of consciousness in some traditional sense (the Kantian unity of apperception, say). Instead, her argument is directed against the notion that human behaviour is centrally controlled, perhaps governed hierarchically. Control, the data show, can fragment in various ways. From here, she draws a very strong inference about traditional conceptions of the self and about the unity of consciousness: both, however the tradition has it, are illusory, not in existence. The premise connecting the data and her conclusion seems to be this: 'If our conception of control changes, then our conception of the self changes *pari passu*.' (Churchland, 1983, 85)

This move seems like a mistake, for three reasons. First, the argument goes through only if, once again, the conception of unity or self targeted is a straw man. Consider the everyday example Churchland cites, doing several things at

once at a party. What conception of unity or self could this count against? The only conception that is clearly contradicted by this is the conception of a self or conscious unity able to do just one thing at a time. Such a thing is a kind of serial processor, able to accomplish only a single task, or control a single action, before it is able to move on to the next. So far as I know, no one is committed to such a thing.

If something like this is fundamental to the conception of consciousness under consideration, then no doubt a self or unity able to do several things at once counts as evidence against it, much like post-combustion weight gain counts as evidence against phlogiston. However, it seems likely that other, better conceptions of consciousness are available that do not require anything like this. It is difficult to imagine what a conception of consciousness would be without some notion of self -- perhaps suitably watered down by Humean considerations -- some subject for whom the contents of consciousness are objects. Certainly, some unity must attend the co-conscious mental states of every conscious subject, properly so called. Nevertheless, the notion of a serial self that Churchland undermines does not seem to be fundamental to every conception of consciousness.

Second, nearly all the data cited are psychopathological cases, and these are murky waters. It is worth at least expressing a certain kind of reservation here. Such cases

might be interpreted as instances in which damage to a brain reveals an underlying disunity that is there all the time, or they might be interpreted as instances in which damage to a brain results in disunity that is never there in normal cases. There are other, far more complex worries as well. We do not know enough about the workings of the brain to interpret such cases with great confidence or design experimental situations that issue in knock-down results. In a recent discussion of the relevance of split-brain studies to philosophical conceptions of the unity of consciousness (Marcel, 1996), a researcher cited elsewhere by Churchland (Churchland, 1992) is at pains to show how difficult it is to arrive at acceptable data from split-brain patients that have any bearing at all on philosophical questions. This is not to say that experimental results have no place in our theorizing, but work with damaged brains is fairly new territory. It is worth remembering when trying to make sense of the implications of any psychopathological case that a number of interpretations, with and without philosophical relevance, are often on more or less equal footing.

Third, the psychopathological data cited might, on some interpretation, unsettle the idea that there is only one unity of consciousness per person or only one self per person, but it does not unsettle the view that any fact of consciousness is in some unity or that conscious events involve a subject of consciousness. Facts about the control of some behaviour, in

other words, seem separable from the nature of the self and the unity of consciousness. Churchland gives us no reason to doubt this.

Third aspect of the modification: the special relation between language and consciousness

The final aspect of the modification is more obscure than the rest, because, as Churchland says, it is difficult to find a substantive expression of it. This should make us suspicious immediately of her claim that allegiance to this dogma is 'widespread, and may perhaps be granted the status of conventional wisdom.' (Churchland, 1983, 85) One would have expected so popular a position to find expression somewhere. Nevertheless, Churchland identifies two different versions of the thesis:

> a strong version: '...language is essential for consciousness...' (Churchland, 1983, 87)
>
> and a weaker version: 'nonverbs are merely conscious in contrast to the verbals, who have *self-consciousness*.' (Churchland, 1983, 87)

Two questions arise with respect to each version. First, what is the nature of the special relationship thought to hold between consciousness and language? Second, is the version in

question really widespread, or is this another straw man?

According to the first version, the relationship is thought to be an essential one. It is tempting here to be reminded of Descartes and try to interpret the first version of the thesis in Cartesian terms. However, Descartes' claims are not what Churchland has in mind. She says that language, according to the first version, 'causes, creates, or otherwise brings about consciousness'. (Churchland, 1983, 87) This is clearly not what Descartes argues in Discourse 5. There his claim just is that language use marks us off from other animals. Even the most dull-witted human is able to respond in appropriate ways to anything said to him, and this, for Descartes, is a sign of reason. Animals, by contrast, lack this ability, and this suggests to Descartes that animal reactions to stimuli are just that, automated reactions, reflexes. Descartes certainly does not claim that language causes or creates consciousness. For him, and probably for most of us, that gets things the wrong way around.

How then are we to understand the claim that language causes or creates consciousness? We are offered the claims of two alleged proponents of the view by way of explanation. Churchland credits Eccles with the belief that 'the non-physical mind makes contact with the brain essentially through the area of the brain involved with language'. (Churchland, 1983, 87) She also maintains that Popper 'argues that consciousness is an emergent property and language is somehow

the vehicle of emergence'. (Churchland, 1983, 87) The two claims seem both incredible and, worrying for Churchland, not obvious expressions of the first version of the thesis under consideration. The fact that they are incredible, and certainly not the sort of thing that might be 'granted the status of conventional wisdom', (Churchland, 1983, 86) suggests that Churchland is setting up a straw man -- it seems likely that, much like the dubious conceptions of global transparency and the serial self, the view that language creates consciousness is not held by many in the history of philosophy.

Proponents of the second version draw 'a distinction which will accord a lower, brutish grade of consciousness to nonverbal animals, and will accord a high grade or full-blooded consciousness to humans.' (Churchland, 1983, 86) The idea seems to be that an entity has the kinds of complex representations we have only if it has a linguistic medium for them. Without such a medium the sort of consciousness a dog has, say, is composed entirely of raw feels (hunger pangs and the like) but no propositional attitudes or self-reflective states ('I am feeling hungry' or, perhaps, 'I would like one of those biscuits').

The evidence against both versions of the targeted thesis comes from certain considerations based on the behaviour of nonverbal creatures and their physiology (13, 14). The nature of some animal behaviour, Churchland argues, requires

crediting them with a system of internal representations and even a conception of self. The behaviour of chimpanzees in front of mirrors, for example, suggests that they have some conception of self that might not be different in kind from our own. Evolutionary and structural parallels serve to strengthen the case, as does a consideration of the fact that we credit preverbal children and nonverbal adults with consciousness.

If the conception of consciousness under consideration requires that only language users are capable of consciousness, or, to put it differently, that language use is a prerequisite for either consciousness generally or the kind of consciousness that we have, then all of this probably does count as evidence against it. However, once again, Churchland gives us no argument for the claim that every conception of consciousness requires this kind of commitment to a strong relation between language and consciousness. It seems at least reasonable to suppose that some conception of consciousness is possible that does not issue in the claim that, say, preverbal children are not conscious. It appears that, once again, Churchland is foisting an unnecessary burden on the concept of consciousness and then attacking the concept for being unnecessarily burdened.

So we now have answers to the two questions raised at the beginning of this section: how are we to understand Churchland's conclusion and what conception of consciousness

is operative in her thinking? We know that her conclusion is that consciousness, like phlogiston before it, faces empirical disproof. We have also seen that the target of Churchland's paper is best understood as a modification of the ordinary conception of consciousness, and we have considered its three aspects in some detail. In addition, we also have a reasonable understanding of the analogy she sets up between the concept of consciousness and dubious concepts like phlogiston.

1.3 Objections and conclusions

Consider this condensed version of Churchland's argument:

(i) There is an 'orthodox conception of consciousness', and its aspects 'include the alleged transparency of the mental, the supposed unity of consciousness and the idea of the self, and the allegedly special relation thought to obtain between language and consciousness.' (Churchland, 1983, 80)

(ii) Certain data emerging from the brain and behavioural sciences suggest that the orthodox conception faces empirical disproof.

(iii) If (ii), then the orthodox conception may be replaced by new conceptions with, as Churchland says, quite different dimensions and dynamics. 'Consciousness' might be doomed to join other undistinguished conceptions like 'phlogiston', 'ether' and the rest and play no part in our understanding of ourselves.

Let us work through the difficulties with the argument, beginning with (i). According to (i), there is an orthodox conception of consciousness. As we saw in section 1.2, it is unlikely that there really is an orthodox conception of consciousness. There is certainly no orthodox body in a position to fix the meaning of the term 'consciousness', as there was in the other cases she cites: spirits (alchemy), crystal spheres (pre-Galilean astronomy), demonic possession (medieval medicine), and so on. Instead, there are a number of different specialized fields with different agendas and different understandings of 'consciousness'. There are workers in each field who typically employ very different conceptions of consciousness, even amongst themselves. Identifying and attacking *the* concept of consciousness does not make much sense in such a context. There just is no single, dominant, orthodox conception.

Using Kuhnian language is helpful here. By Churchland's lights, there is a consciousness-paradigm, and, given the data she cites, it might be in crisis, on the verge of revolution. Anomalous data are piling up, threatening a paradigm shift. However, it seems likely that our situation is something closer to pre-science, that is to say that we have not yet arrived at anything like a paradigm in our efforts to understand consciousness. We do not have a set of widely accepted generalizations about consciousness and are not using them in conjunction with a certain set of techniques, as Kuhn

might put it, to engage in puzzle-solving activity.

If the authors of a recent collection, *Explaining Consciousness* (Shear, 1997), are representative, there is not even agreement about which puzzles are the most pressing or how we might begin to solve them. Some, like Chalmers, identify 'the hard problem' of consciousness as accounting for the existence of experience in a material world, contrasting this with 'the easy problems' of understanding the neural correlates and functional properties of cognitive processes. Others claim that Chalmers' distinction between hard and easy problems of consciousness is ill-founded: *all* the problems are hard. Still others deny that there is a hard problem (or problems) in the first place. In the same book, some thinkers argue that the hard problem cannot be solved while others maintain that it already is solved. Competing methodologies abound.

With respect to consciousness, a very good case can be made for the claim that we are not yet in normal science. Compared to the sort of work that might go on under the structured direction of a paradigm, contemporary philosophers and other researchers are engaged the kind of disorganized and diverse activity characteristic of pre-science: arguments over fundamental definitions of subject matter and the identification of important problems, disagreements so widespread that individuals are forced begin with their own definitions and justifications of the assumptions that they

make, before they can even begin to address problems that they themselves have to identify in the first place. In this climate, an attack on *the* concept of consciousness is wrongheaded.

If there is no orthodox conception of consciousness, then Churchland might be understood as attacking either the ordinary conception of consciousness or some common modification of it. There are, as we saw, good reasons for understanding her as targeting a modification. Part of the burden of section two of this chapter was to show that, if it is a modification under attack, it is certainly not a widely held one. It would have been of interest if Churchland identified assumptions common to most or even many contemporary conceptions of consciousness and demonstrated that those assumptions are in conflict with data. However, when the three aspects are brought out into the light, it looks as though Churchland is setting up and knocking down a straw man. It is not clear that anyone is a proponent of the modification of the conception of consciousness that Churchland attacks, and it is therefore difficult to see the relevance of her conclusion.

Though it seems unlikely that she is attacking the ordinary conception directly, she might be attacking it in a roundabout way by arguing that the modification identified is problematic because of its roots in the ordinary conception. It might be possible to argue that any modification of the

ordinary conception is doomed because the ordinary conception itself is deeply flawed. The modification she identifies is just a particularly good example of how bad a conception of consciousness can be if the original flaws in the ordinary conception are brought on board, or so Churchland might claim.

Reflection on the everyday conception of consciousness and the three aspects Churchland identifies just does not bear this out. Examples will help make the point. In the ordinary sense, we introspect our conscious states regularly. Consider the commonplace of a smoker lighting up and reflecting on his habit: He asks himself why he continually smokes, despite knowing the risks to his health and the cost; he spends a moment weighing his desire to quit with the pleasant feelings associated with the first smoke of the day, and so on. He may identify his motivations, or possibly come to the conclusion that he has no idea why he smokes. Introspection, on the ordinary view, is, among other things, examining one's thoughts, feelings, perceptions and the like very closely. So far as I can tell, there is no commitment here to the view that in the act of introspection all one's mental life, in Churchland's broad sense, is laid bare to the thinker. This much is, at least, clear to those who smoke but fail to understand why they do so.

The smoker might (as many smokers do) discover on lighting up that he already has a cigarette burning in the ashtray. Not only might he admit that he has no idea why he

lights up in general, but, in this particular instance, it becomes clear to him that sometimes has no idea even when he is smoking. These admissions are not anomalies, but familiar parts of our inner lives and are probably built into our ordinary conception of consciousness. The ordinary person knows, in virtue of commonplaces like this, that it is possible to forget oneself, lose oneself in anger, passion or, perhaps in the case of the smoker, concentration. If this is really a part of our ordinary understanding of ourselves, it seems unlikely that the ordinary conception is committed to the strong view that there is a supervenient self in charge of all of our actions. To be sure, the ordinary conception makes some watered-down commitment to the idea of an 'I' in charge of at least some actions. However, the fact that everyday experiences of 'cognos' -- mental typos like having two cigarettes burning in the ashtray -- are everyday experiences suggests that they are not stark anomalies for the ordinary conception, but part of it.

No doubt the ordinary conception places some emphasis on the relation between thought and language -- we sometimes tell whether or not a person is conscious by asking them. However, it is simply not the case that the ordinary conception confines the class of conscious (or self-conscious) things to users of language. It would be difficult to find an everyday person willing to claim that babies and nonverbal adults are not conscious.

The problem here is that the ordinary conception makes some rough and ready commitment to introspection, the self, and a relation between language and consciousness, but in each case the commitment is innocuous and extremely loose. People can ascertain their own thoughts; I am in control of some of my actions; conscious people can usually talk, and so on. All three of these claims can be understood as truisms associated with the ordinary conception. They can also be pressed into service as general philosophical claims, and this is where care is needed. If they are interpreted unnecessarily strongly, as Churchland interprets them, as commitments to global transparency, a serial self, and an essential relation between language and consciousness, then of course they are in conflict with data. They are also in conflict which much that we know to be true about ourselves in an ordinary sense, as the examples above make plain.

The point is that the ordinary conception does not characterize introspection, the self, and language this way, and it seems likely that more careful interpretations will deal better with the data Churchland adduces. In the end, it is clear that the fate of the unnecessarily strong modification Churchland attacks has little bearing on the fate of the ordinary conception or more careful modifications of it.

We have come to several conclusions about the first part of Churchland's argument, (i). First, there just is no

orthodox conception of consciousness, and this fact renders her general project suspect. Second, a careful consideration of the concept she attacks suggests that she is setting up and knocking down a straw man. It is unlikely that anyone actually holds the three theses identified, much less that they have the status of conventional wisdom in contemporary philosophy of mind. Third, Churchland provides no reason for the belief that problems with the modification identified suggest deeper difficulties with the ordinary conception of consciousness.

(ii) and (iii) can be considered much more briefly. According to (ii) the modification under consideration faces empirical disproof. As we have seen, the fate of the modification Churchland identifies seems more or less irrelevant to the ordinary conception and better modifications of it. The conception Churchland attacks makes unnecessarily strong claims that are not really a part of the ordinary conception and certainly no essential part of any conception of consciousness as such. According to (iii), the concept of consciousness might go the way of phlogiston and other dubious concepts. This, of course, only follows if (i) and (ii) are not problematic, and there are reasons to think that they are. Nevertheless, it is worth noting how weak the analogy between consciousness and concepts like phlogiston actually is. Phlogiston had a clear place in a theory of combustion, with well articulated aspects that came into conflict with data.

However, there is no orthodox conception of consciousness with clearly defined aspects that might serve as the target for empirical disproof -- hopefully, some conception will be good enough to come to the fore, but that does not seem to have happened yet. Moreover, phlogiston theorists encountered data that should have been impossible, experimental results that contradicted one of their fundamental assumptions about the nature of burning. The data Churchland cites might well stand in conflict with the dubious modification she identifies, but, as we have seen, not much follows from this about the everyday conception or about better modifications of it.

It is also worth emphasizing the fact that anomalous data encountered by phlogiston theorists should have been impossible from their point of view. The data Churchland cites, puzzling though they are, do not strike us in the same way. If the data cited did strike us as the discovery of post-combustion weight struck phlogiston theorists -- if we really saw the data as impossible and we really did operate with the conception of consciousness Churchland credits us with -- the result would be earth shattering. We would have reason to believe that a huge chunk of our beliefs about ourselves and others is simply false. The fact that the data do not have this effect on us suggests either that Churchland is attacking the wrong conception of consciousness or that the data really are not analogous to those encountered by phlogiston theorists -- perhaps both possibilities are true.

At any rate, if there is an argument by analogy in Churchland's thinking, its force is considerably weakened by all of these considerations.

All of this is not to say that empirical results have no place in our understanding of consciousness or mental phenomena generally. Philosophizing about the nature of consciousness has to take into account results in the sciences; the philosophy of mind needs all the help it can get. However, the particular tactic Churchland adopts, undermining *the* concept of consciousness with recalcitrant data, seems the wrong tactic, given the state of our understanding of consciousness and the brain. Instead of using data in an attack on some conception of consciousness, perhaps a better tactic is trying to develop a conception of consciousness in the context of data, a tactic Churchland also supports. (Churchland, 1983, 95) In the end, it seems clear that despite Churchland's claims it is still reasonable to hope that some conception of consciousness will figure into our understanding of the mind; it is still reasonable to hope, that is, that consciousness is a thing that can be understood.

2. Evidence for the Existence of Consciousness

It may be true that the attempt to undermine *the* concept of consciousness with empirical results is misguided. However, to turn the problem on its head, what evidence is there for crediting ourselves with conscious states in the first place? In the last chapter, a number of references were made to the everyday or ordinary conception of consciousness. What is the source of our confidence that we are conscious in this everyday sense? It might be said that we know we are conscious simply because we are conscious; we know we are conscious because we experience certain properties directly. But there is an obvious circularity in this response, however reasonable it might seem in some contexts. Can we do better than this?

Rey formulates a unique challenge to the presumably innocuous view that we are all conscious in a certain sense, and a consideration of his challenge might issue in the evidence we are after. He begins by following Levine (1983, 1988) and others in identifying an apparent explanatory gap between physical and conscious phenomena. In a great many cases, Rey says, we manage to explain a macro phenomenon by giving a theoretical account of a certain micro phenomenon or phenomena. This account shows how micro phenomena upwardly necessitate macro phenomena, that is, how the existence of macro properties are a necessary consequence of the truth of a micro theory, in conjunction with auxiliary hypotheses and

scientific laws.

For example, the expansion of water when it freezes is a necessary consequence of the laws of physics and facts about the bonding properties of water molecules: the lattice structure it assumes at temperatures below 32 degrees Fahrenheit takes up more space. However, we seem unable to say how any facts about the brain could necessitate something's being a conscious state in the way that we are able to say how facts about the nature of water necessitate its expansion at certain temperatures. As Rey says,

> It seems that when we reflect upon our concepts of (to use the jargon for 'the way things feel') 'qualia' and consciousness, no explanation in physical, or even physical/computational terms seems available. (Rey, 1995, 128)

There is, then, an explanatory gap between conscious phenomena and what we know about the brain. Nothing we can say about the brain issues in an explanation of consciousness in terms of upward necessitation, or so Rey maintains.

Rey characterizes his unique solution to -- or, rather, dissolution of -- the problem posed by the explanatory gap with reference to an imaginary computer, which he calls 'the COG computer'. The computer itself is a relative of other, more famous denizens of thought experiments in the philosophy

of mind. It is, in many ways, like Block's (1991) homunculi-headed robot -- a machine with a hollow head filled with little men who by concerted efforts manage to realize the same functional description as you or me. Block's point is, of course, that it would be a mistake to credit the robot with mental states, in particular qualitative states, yet, he claims, that at least some versions of functionalism are forced to do just that. Block's conclusion is that functionalism is false.

Rey imagines a computer that realizes the same functional description that a given person might, and agrees that there is a sense in which it would be a mistake to credit it with certain conscious properties. However, instead of agreeing with Block that functionalism does not secure an adequate characterization of mental states, Rey maintains that reflection on the computer shows that there is no good evidence for the existence of certain conscious properties in the first place -- in the computer or in us. It is a mistake to credit things like a homunculi-headed robot with a certain sort of inner life, and it is a mistake for us to credit ourselves with a certain sort of inner life too.

If Rey is right, then in a sense the claim that consciousness can be understood is much mistaken. Part of what it means to say that consciousness can be understood, as that claim is being used here, is that there really is something there to be understood in the first place, and the

relationship between that something and the brain can be understood as well. However, if Rey's arguments go through, there is no explanatory gap to close at all -- phenomena on one side of the gap, conscious properties as we usually understand them -- do not really exist.

We will deal with this claim by beginning with a recapitulation of Rey's description of the COG computer. In section two, we will consider the distinction that Rey draws between the sense in which the computer is conscious and the sense in which we claim to be conscious, what Rey calls 'weak and strong conceptions of consciousness'. In section three, we will examine the nature of introspective evidence for the self-ascription of strong conscious properties. Finally, the last section is an evaluation of Rey's claims concerning the evidence that one might have for the self-ascription of strong conscious properties, and there I will try to set forth an argument for the existence of consciousness in the strong sense. I hope to show that Rey's dissolution does not work; the problem of understanding the relation between what we normally think of as conscious states and the brain really is a problem. We begin, then, with the COG computer.

2.1 The COG Computer Thought Experiment

Rey imagines a computer that realizes certain aspects of human mental processing based on some of the computational accounts postulated by cognitive scientists and

functionalists. It is easier than one might think, he maintains, to program a computer replete with,

> ...perceptual states, intentional states, intentional states about intentional states (thoughts about thoughts about thoughts...), "self-consciousness", attention, planning, decision-making, use of a natural language.... (Rey, 1995, 128)

His argument for this claim is based at least in part on a version of the representational theory of mind.

Representation, rationality intentionality and COG

According to Rey, if anything justifies the ascription of propositional attitudes to a creature, in particular beliefs and preferences, it is the satisfaction of what he calls 'rational regularities': 'regularities among a creature's states whereby it instantiates the steps of inductive...deductive...and practical reasoning.' (Rey, 1983, 7) Animals, for example, seem able to act in ways that lead to the acquisition of certain desired objects in their environment, and they do so reasonably. That is to say, they seem to try to get what they want in a manner that accords with inductive, deductive and practical reasoning. The best explanation for this sort of behavior, Rey argues, is that

mental processes of some kind are going on inside the animals: they have beliefs and desires coordinated by induction, deduction, and means-ends reasoning. That is the best, perhaps the only, reason one might have for ascribing mental states to a system, according to Rey.

He claims that, in general, the best reason one might have for ascribing mental states to someone or something is by observing behavior, looking for patterns in the way a thing behaves that belie an underlying inner life. If what we are looking for is evidence for mental states, adherence to the rational regularities is the best evidence we can have.

In addition to rationality, systems with mental states also exhibit intentionality. Rey understands intentionality according to a version of Brentano's formula: mental states are intentional insofar as they seem to involve a relation 'not to an existent thing preferred or believed about, but to something else -- an idea, a possibility, a proposition -- in any case, a *representation* of something that may or may not turn out to be real'. (Rey, 1983, 8) Rey moves very quickly here, and it is not clear how we are supposed to take the claim that the COG computer exhibits intentionality. This is a worry we will return to momentarily.

These two features, rationality and intentionality, are distinctive of mentality, and, according to Rey, the best theories currently available that take account of both are syntactic, i.e. they deal with the formal properties of

representations. It can be shown, he points out, that all deductively valid arguments in a given domain can be generated by rules defined over objects that are entirely syntactically specified. (Henkin, 1949, 1969; Kripke, 1959) He argues that this gives us at least some ground for hoping that other kinds of reasoning might be understood syntactically as well. Citing Fodor's (1975) arguments, Rey maintains that a formal language, that is, a syntactically specified set of sentences encoded in the brain, seems like the right kind of thing to do the job of mental representation. Rey concludes:

> "thinking" can be increasingly characterized as a very particular kind of computational process, namely, one that involves *syntactic transformations of representations...thinking is spelling*. (Rey, 1983, 9)

According to this view, the mind is essentially a computational engine, building up a system of representations in a language of thought via computational processes, or, as some say, symbol or number crunching. Thinking, on this view, just is the manipulation of inner symbols according to a set of rules. So Rey claims that 'any system that could consistently spell and transform strings of symbols in accordance with particular rules would qualify thereby as a thinking thing.' (Rey, 1983, 9)

On hearing this claim, one immediately wonders what happened to the intentional properties of the computer. There are a number of very well known objections to the line Rey is taking here. Searle (1984), in perhaps the best known example of an objection to this kind of view, imagines an English speaker who knows no Chinese locked in a room. Inside the room are two batches of Chinese writing and instructions in English for correlating them. Individuals on the outside send in questions written in Chinese. The person on the inside matches these symbols, based on their formal properties alone, to the first batch, correlates them with the second batch according to the rules, and passes the results from the second batch back out of the room.

Suppose that the rules are set up such that the resulting symbols are the sorts of things a Chinese speaker might say in response to the questions asked. From the outside, it looks like real understanding is going on in the room, but since by hypothesis the person in the room speaks no Chinese, all that is really going on is the manipulation of symbols based on their formal properties. This is, according to Searle, all that a computer does: the correlation of input with output according to a set of rules. Thinking is something more than syntactic transformations, on this view. Those sympathetic to Searle's view that nothing can think purely in virtue of manipulating symbols are going to have considerable difficulty accepting Rey's claim that the COG computer is a thinking thing.

For the moment, let us bracket this and other objections and grant Rey a great deal in order to pursue his claims about introspection and the evidence one might have for the existence of conscious properties. It is, at any rate, possible to raise the kind of worries that Rey does without invoking thinking computers, and more will be said about this in the next section. So we continue with Rey's characterization of the computer.

Consciousness and the COG computer's program

Suppose that the computer has been programmed such that it is able to realize particular aspects of human psychology. Its program is characterized by eight clauses; here are the first five, in Rey's own words:

1. The alphabet, formation and transformation rules for quantified modal logic (the system's "language of thought").
2. The axioms of your favorite inductive logic and/or abductive system of hypothesis, with a "reasonable" function for selecting among them on the basis of given input.
3. The axioms of your favorite decision theory, and some set of basic preferences.
4. Mechanical inputs, via sensory transducers, for Clauses 2 and 3.

5. Mechanical connections that permit the machine to realize its outputs (e.g., its "most preferred" basic act descriptions). (Rey, 1983, 10)

Rey contends that any computer functioning according to the program here described would both satisfy the rational regularities and be a full-blooded intentional system.

The computer he describes is hooked up to the world via sensory transducers that enable it to have what Rey claims amount to perceptual states. It has other sorts of connections too that facilitate the emission of various sorts of output -- it has a speech synthesizer that allows it to emit the sounds attending a natural language and robotic arms that enable it to manipulate things in its environment, and so in a sense behave. It also has a recursive subroutine connected to its buffer memory that, according to Rey, enables it to introspect the processes going on in some parts of its sub-modules. Some of its axioms include elements of decision theory, and for Rey this means that it can make plans or focus its perceptual apparatus on some particular part of the environment or other. In such ways, Rey maintains, the COG computer might realize the sorts of states humans realize when they think, believe, introspect, speak, and so on.

In what sense is the computer thinking, believing, introspecting, and speaking? It is clear that all of these capacities are to be understood in broadly functionalist

terms. As we have seen, Rey maintains that thinking and believing can be characterized as a kind of computational process: the manipulation of symbols according to their formal properties. It is clear that he is taking up computationalism and a functionalist view of mental states. On his view at least some of the states that the computer enters into are beliefs and desires because those states stand in the appropriate functionally defined relationships to input, output and other inner states.

A speech act for Rey must be something more than just uttering the characteristic sounds of a natural language: otherwise parrots and tape recorders can speak. For a thing to be credited with the capacity to speak, it must be able to pass through the right sorts of inner states, related in the right way to other inner states and caused by the right sort of stimuli -- the same sorts of inner states that a normal person goes through when, for example, replying to another in conversation.

What makes the inner states, stimuli and output the right ones is not made clear. However, if thinking and believing really are just a matter of functionally specified symbol manipulation, then by hypothesis, the computer can realize the same states a person does when he thinks, 'I need a drink'. Making the computer capable of speech, for Rey, is just a matter of ensuring that the computer's belief generator is hooked up in the right way to its speech centre, and this, as

Rey describes it, is the case for the COG computer.

What conception of introspection is operative in Rey's thinking? He adopts a computer-model view of introspection. It comes as no surprise, then, that he credits the COG computer with the capacity to introspect. Perhaps the easiest way to understand this model is in terms of machine functionalism. According to the computer-model, a computer's machine table might include the instruction, 'Print "I am in state S" when in state S'. This results in the output, 'I am in state S' whenever the computer is in state S.

Some find this model attractive, because it dispenses with some of the more dubious trappings of some accounts of introspection. There is no suspect Cartesian inner theatre here, nor is there a special organ of introspection analogous to sense organs in the case of ordinary, outer perception -- an organ critics argue cannot be found. Further, according to the computer-model, introspective reports are generated just because a subject is in a certain state, and no special faculty or process of ascertaining the nature of one's inner states is needed. This way of thinking of introspection has the added advantage of avoiding Comte's objection to other models which he alleges presuppose an impossible split in consciousness between the object observed and the observer.

The computer-model, though, is generally viewed as an analogue to what is going on when humans introspect, but Rey seems to take it as an actual description of introspection,

and this is suggestive. Perhaps the most telling response to this model is that it might give a good account of what is going on when a computer prints or otherwise 'reports' its inner states, but it is not clear at all that something like this is going on when humans introspect. Much of the motivation for this response comes from reflection on the experience of introspection itself, arguably something more than the kind of if-then loop posited by the computer model. At least, it might be said, it seems that way from the inside.

The fact that Rey does not take the model as a model, but as a full-blown description of introspection -- such that he credits the COG computer with the ability to introspect because its program issues in printed or vocalized reports of data manipulated -- suggests that he is not swayed by reflection on the experience of introspection. This aspect of his thinking becomes even more manifest in his later claims about the evidence one might have for the existence of certain conscious properties, which we will come to in section three.

Though Rey is willing to characterize the computer as a thinking thing, he refrains from crediting it with consciousness. Thinking, for him, is a matter of symbol manipulation according to rules, and the COG computer does that well enough. However, he says, the program is not 'sophisticated' enough for us to consider it conscious. Further, he claims that there are other systems that satisfy clauses 1-5 that are not normally called 'conscious'. For

example, the subconscious processors in us that subserve perception and language might be called thinking systems, on his view, but not conscious ones. What more is needed? He considers several possibilities, adding several more clauses to the five listed so far.

The first concerns a kind of sophistication of the computer's believing capacity. A first-order intentional system is just a system that satisfies the rational regularities and so has beliefs and preferences as Rey understands them -- for example, the COG computer, if Rey is right about it. A second-order intentional system is a first-order intentional system that has beliefs and preferences about beliefs and preferences. It might be claimed that consciousness consists in having a certain depth of nested beliefs and preferences about other beliefs and preferences.

However, Rey maintains that it would be a simple matter to add the following clause to the computer's program:

6. The recursive believer system.

The recursive believer system is just a subroutine that extrapolates motives, beliefs, preferences and so on from a given set of data by means of iterations of its own means-ends reasoning system. A program with such a subroutine is already in existence, he says, and it is able to extrapolate the motives, beliefs and preferences of characters in a given

story. (Schmidt and D'Addami, 1973; Brown, 1974) The COG computer, equipped with such a system, would then count as n-order intentional for as many nestings of beliefs and preferences about beliefs and preferences as would be required.

However, this would not render the system conscious, according to Rey, because the mere presence of nested reasoning envisaged above does not entail that the system is actually consciously entertaining those nested beliefs. Further, he says, this suggests that requiring the capacity for self-consciousness -- which is, for Rey, just a kind of n-order intentionality about the subject's own beliefs and preferences -- will not secure consciousness either. Even if the computer could refer to itself in its own inner machinations, he argues, it is doubtful that this alone would render it conscious.

Given the sorts of conclusions concerning the presence or absence of consciousness that Rey is making here, it might be hoped that we could determine just what he means by 'consciousness'. If he says, as he often does, that the computer is not conscious just because we have added a certain subroutine, for example, he must know what consciousness is, and perhaps we can extrapolate something from his claims about this conception of consciousness. Unfortunately, he seems to be just straightforwardly denying conscious states to the computer, with little or no explanation. 'Of course', he

seems to be saying, 'no one would think that just the addition of a subroutine would render the computer conscious'. Not much more about his conception of consciousness is forthcoming, but we will be in a position to characterize his view in section two.

What about the capacity to report on the nature of one's own mental states; would this ability render the computer conscious? This, for Rey, would again be an easy matter of adding to the program instructions characterized by the following clause:

> 7. A fragment of English adequate to describe or express the mental states entered in executing Clauses 1-6, descriptions that are produced as a reliable consequence of being in those states.(Rey, 1983, 16)

Whenever an introspective report is required, information from the system's buffer memory is translated into the appropriate bit of English, or perhaps whenever certain states are entered, the computer just prints or otherwise expresses something appropriate about those states, just as a consequence of being in them.

Would this render the system conscious? Again, Rey thinks not. He argues that if clauses 1-6 were not enough to make the system a conscious one, why think that adding a

special purpose compiler like the one envisaged would make much difference?

He next considers sensations, and maintains that although the machine already has the functional equivalent of sensations, most of us would object to the claim that the computer really has experiences, like we do when we see red, for example. Its sensory transducers emit characteristic outputs whenever red is detected, and the usual causal consequences that attend seeing red in us obtain in the computer too. So what more, Rey asks, is there to the sensation of red? Why think there is something more to it in our own case?

The response, that we have some kind of special access to our experiences, an access that gives us reason to believe that there is more to seeing red than standing in certain causal relations, is treated very briefly by Rey at this stage of the argument -- probably because he considers it more carefully when arguing against introspective evidence for a certain understanding of conscious properties. His claim here, though, is that the response seems to presuppose that believing one has a sensation entails one's having it. If that is true, the COG computer is sophisticated enough to acquire sensory beliefs, including the belief that it seems to see red, and that belief is hard to distinguish, if it can be distinguished, from the belief that it is having a red sensation.

If it is claimed that something more than just this belief is going on in our own case, say privileged access or direct acquaintance with a certain kind of conscious property, then Rey will press for evidence that we, in fact, have privileged access or direct acquaintance with anything. He stops here, reserving a more detailed consideration of the nature of evidence for certain properties of consciousness for a later section, which we will come to in a moment.

His point, though, is that if having sensations is required for consciousness, then something more than just the belief, 'I am having the sensation associated with seeing red', is needed. The computer can have that belief too, Rey argues. If it is claimed that, in our case but not in the computer's case, we have a special kind of access to the sensations that give rise to the belief, 'I am having the sensation of red,' then evidence is needed for the existence of the access, and it is not clear that evidence is available, as we shall see. The bottom line is that both the computer and a person can share the same sorts of beliefs about sensations, so those beliefs must not be enough to secure consciousness, or so Rey maintains. After all, the COG computer might have such beliefs, but those beliefs alone do not seem to be enough to warrant crediting it with consciousness.

No doubt it seems very much to us that we are conscious, Rey admits. We have what he calls 'the Cartesian Intuition',

a network of beliefs about our conscious lives and the presumed fact that we are not just programmed artifacts. However, the computer could be programmed with the additional clause:

> 8. The Cartesian Intuition and related claims of epistemological privilege. (Rey, 1983, 22)

Underlying our certainty that we are conscious and the computer is not, Rey argues, is largely just a number of relatively trivial second-order beliefs, such as 'I see clearly that there is nothing which is easier for me to know than my own mind', and 'no matter what your theory and instruments might say, they can never give me reason to think that I am not conscious right here, now.' (Rey, 1983, 22) Rey maintains that this is not enough for consciousness, because the computer could be programmed to have the Cartesian Intuition too.

2.2 Strong and Weak Conceptions of Consciousness

Once this picture of the computer is in place, Rey observes that most people would not consider the computer conscious and asks why. So far Rey has found no reason to credit the computer with consciousness, but in working through the possibilities, it seems that there is not much reason to credit ourselves with consciousness either. All of the

abilities and properties that one might point to in one's own case as grounds for maintaining that one is conscious might be satisfied by the computer, but most of us are not willing to credit it with consciousness. Why the confidence in our own case?

Rey claims that most people maintain that 'conscious states are not *the kind of states* that will be captured by *any* such computational program'. (Rey, 1995, 129) Some claim that an additional dualistic property is needed, while others maintain that some unspecified physical property is required. Rey follows Dennett (1991) in calling such people 'qualiaphiles' and groups their views together under the heading ' COG-transcendent conceptions of consciousness' or 'strong notions of consciousness'.

These views of consciousness are to be distinguished from weak notions, which involve the sorts of properties that he maintains could be exhibited by the COG computer: wakefulness, attention, and introspectibility. When the computer is switched on and stands ready to process information, it is wakeful. When its perceptual apparatus is directed toward some object in its environment and its inner states stand in the right relations to that thing in the world, its outputs and other inner states, it can be said to be attending to it. As we have seen, when it prints out a report which records something about one or another of its inner states in virtue of commands written into its

programming, it can be said to introspect. This, and not much more, is all Rey tells us about weak consciousness. What more might be extrapolated about his conceptions of strong and weak consciousness?

Since he takes strong conscious properties to be phenomena lying on one side of the explanatory gap, reconsidering his characterization of the gap is useful. He says,

> ...what physical facts could possibility *necessitate* something's *looking green*, or *red* or being a conscious state at all? It seems that when we reflect upon our concepts of (to use the jargon for 'the way things feel') 'qualia' and consciousness, no explanation in physical, or even physical/computational terms seems available. (Rey, 1995, 128)

It sounds as though, for Rey, 'something's looking green or red' are examples of qualia, and 'qualia' is as he says just jargon for the way things feel. The language is confusing because, presumably, the qualitative properties being denied are not some thing's -- properties had by some thing out there in the world -- but properties of some conscious states.

Rey's claims, it seems obvious, are not about properties in the external world, but about the existence of properties

of certain conscious states. The explanatory gap exists between what we know about the brain and what we think we know about consciousness, on Rey's view, so it has nothing to do with the properties that material objects in general might have, at least not in any direct way. So though Rey says that what does not exist is 'something's looking green', he must not be denying the existence of a property had by some thing out there, but allegedly had by at least some conscious states. It is in virtue of those states, some suppose, that things look green to us, or more generally, that things feel a certain way to us. It is this sense of 'the way things seem' that the physical and computational facts seem unable to explain.

This fits in with what Rey says about the kind of states some suppose are not captured by the COG program. He says that defenders of strong conscious properties maintain that there is something more, some dualistic or unspecified physical property that the computer lacks. That property or set of properties is, as Rey claims, cognitively transcendent, something not captured by the purely computational specifications of the computer.

Compare this conception of consciousness to what Rey calls the weak conception, the kind of conscious properties that the COG computer has: 'mere wakefulness, attention and introspectibility'. (Rey, 1995, 130) If the COG computer has weak but not strong consciousness, then it must somehow be

able to realize wakeful, attentive and introspective states without strong consciousness. Given what Rey claims about qualia, the computer must be able to realize those states without anything seeming a certain way to it. It must be possible, for example, for the COG computer to focus its attention on a green object, while nevertheless nothing seems green to it. How are we to render this sort of thing intelligible? If we cannot, how are we to render the notions of strong and weak consciousness intelligible?

One way that philosophers soften up our intuitions in this area is with the alleged logical possibility of zombies (Chalmers, 1996, for example) and their functional descriptions. It is possible, some maintain, to imagine a creature that is molecule for molecule identical to you but nevertheless lacks conscious experience. Suppose you are drinking a cold beer and considering the pleasant hoppy aftertaste and cool sensations in your fingertips as you hold the glass. If your zombie twin is located in a sufficiently similar environment and shares a history sufficiently similar to your own, he is enmeshed in the same causal network of sensory inputs, motor outputs, and informational states as you are. That is to say, he is functionally identical to you: he reacts as you do to the same sensory inputs, and information is processed in him as it is in you. In particular, say, he will respond as you do if asked about the beer's taste; if prompted, he will draw his attention to the condensation on

the glass just as you might, and so on.

Judging by his behaviour and his general functional description, he is as awake and attentive to his surroundings as you are, and given his reports, he seems as pleased with the beer as you are. Despite all this, there is nothing of a phenomenal nature accompanying his wakefulness, attention or introspective access. There is nothing more to him than what Rey is calling weak consciousness: his inner states have none of the qualitative feels associated with seeing green, tasting beer, or touching glass. There are, to be sure, informational states inside your zombie, and these stand in a functionally defined relation to other informational states and his behavior, but that is more or less it. Things do not seem a certain way to him, as, some claim, they do to you.

Another way to understand the difference between you and your zombie twin (and between strong and weak consciousness) is by making a distinction between phenomenal and psychological aspects of the mind. Chalmers (1996) understands the distinction in terms of feeling and doing. Phenomenal conceptions of consciousness place emphasis on what it is like to be in a particular mental state, how things seem to a subject, and so on. What role consciousness plays in the causal web is obviously important to such conceptions, but what matters most is not what conscious states do, but how they feel. For example, it is the fact that pain feels a certain way to a subject that makes pain the state that it is.

By contrast, psychological conceptions are concerned with the mind only insofar as it is the causal basis for some behavior. It is what causes pain and pain's ensuing effects that make pain the state that it is -- how it feels to the subject, if it feels any way at all, does not enter into the essence of the state.

So when Rey claims that the COG computer is only weakly conscious, that it realizes wakeful, attentive and introspective states without qualia, he is saying that the computer is conscious in the psychological but not phenomenal sense. Its states do the same things as yours do -- that is, they enter into the same causal relations as do yours -- but it does not feel like anything to be a COG computer, just as it does not feel like anything to be a stone. That is how it is possible for the COG computer to focus its attention on a green object, while nevertheless nothing seems green to it.

Attention, according to the psychological conception, is a matter of standing in the right set of causal relations to, say, a certain patch of green. Because, by hypothesis, the computer is equipped with the relevant transducers and processors, it can enter into the right relations. If that is all there is to weak attention -- and all there is to the COG computer's mental life is weak states of consciousness -- then the COG computer can weakly attend to a green thing, without anything seeming green to it.

It might be said that what Rey calls 'strong consciousness' just is what is normally meant by 'consciousness' itself, and what he calls 'weak consciousness' is something else entirely, something not properly called 'consciousness'. This might lead one to suppose that it is a mistake to admit that the COG computer is conscious in any sense at all. Not much hangs on how the computer's properties are characterized, for Rey's argument requires only that we deny whatever conscious properties we claim to have to the computer, and this most of us are willing to do. His interest, early on, is not to characterize the sense in which we claim to be conscious, but to ask why we deny whatever conscious properties we claim to have to the computer.

2.3 Introspective Evidence for Strong Conscious Properties

It is reflection on this denial, the denial of strong consciousness to the COG computer, and also reflection on the ascription of strong consciousness to ourselves, that leads Rey to the following question:

> ...what reason does one have for supposing in one's own case that one *does in fact* have what it takes, i.e. phenomena that are not necessitated by a mere realization of COG? What does one know about oneself that rules out one's self being essentially a COG computer?...If we...think we are conscious and

> the COG computer isn't, we need to ask what entitles us to our confidence. (Rey, 1995, 130)

The COG computer realizes, by hypothesis, the same computational states that we realize when we perceive, think, plan, introspect and so on, so the difference between us and it cannot be computational. There are no doubt physical differences between us and the computer as well, but it seems extremely risky to claim that a system cannot be conscious unless it is built of the same stuff and in the same way we are. Further, our confidence in the belief that we have strong conscious properties cannot lie in the physical facts about the brain as we know them, according to Rey. As we saw at the outset, he claims that nothing about those facts upwardly necessitates strong conscious properties. So what entitles us to the belief that we have extra properties that COG lacks? What, in other words, gives us reason to ascribe strong conscious states to ourselves? This is the crux of things for Rey, and much of his work is a sustained attempt to undermine the natural answer to this question.

As mentioned earlier, there are a number of reasons to deny that computers can think and stop Rey's arguments further upstream. However, we have been willing to grant a great deal to Rey in order to arrive at the question, what is it in our own case that makes us confident that we have strong conscious properties? It is clear enough that the same question can be

asked whether or not we follow Rey in his claim that all there is to thinking is spelling, or that there is a language of thought, and so on. Even if his claims about computers are thoroughly wrongheaded, as it might be argued they are, the question he asks is still a meaningful and interesting one: what ground do we have for thinking we are conscious in the strong sense?

Introspecting strong conscious properties

The obvious response seems to be that it is somehow in virtue of having strong conscious properties that we have reason for ascribing them to ourselves. In particular, it is our introspective access to conscious states that warrants the ascription of something more than COG consciousness in our own case. We know there is something more to us than computational processes because we have access to the properties in question. Many admit that there is nothing in the physical facts themselves that leads us to conclude that we are conscious -- indeed, it is sometimes said that if we did not know about consciousness directly in our own case, there would not be an explanatory gap in the first place, because we would never have posited the existence of consciousness at all.

This kind of response to the sceptical suspicions underlying Rey's question takes several forms. Chalmers, for example, maintains that 'we know about consciousness more

directly than we know about anything else, so "proof" [for the claim that we are conscious] is inappropriate.' (Chalmers, 1996, xiii) Some might even say that proof is impossible. If asked what grounds we have for thinking that we are conscious, no argument in the traditional sense can be given, because there are no premises more certain than the argument's conclusion. The best we can do, it might be said, is appeal to a kind of direct introspective knowledge that each of us has in virtue of the fact that we instantiate conscious properties. We just know that we are conscious because we are conscious -- there are no mediating inferential steps with which to build an argument.

McGinn puts much the same point in this way,

> ...to deny that one is conscious requires one to deny what is self-evident....conscious states are *data* -- part of what the world presents to us as simply so. (McGinn, 1993, 35)

Self-evident truths are often considered suspect, appealing as they do to the notion that there are some truths that can be apprehended simpliciter. Nevertheless, in this case the idea seems to be that we know we are strongly conscious because strong conscious properties are just obvious features of the world or at least obvious features of the way the world is to us.

From all of this, the following response to Rey's question might be distilled. We know that we have strong conscious properties because we introspect them. This is to claim that we have a kind of direct, which is to say noninferentially mediated, experience of them. There is no proof for the claim that we are strongly conscious, if by proof what is required is an argument, but nevertheless we do have good reason for ascribing strong conscious properties to ourselves: we have direct experience of them in introspection.

Rey's three replies

Rey articulates the following three counter arguments to this move. We will consider each argument in outline first, then take account of the position that Rey occupies in virtue of them. Once all of this is in place, we will be in a position to evaluate the arguments and come to some conclusions about Rey's view of conscious properties.

First of all, he argues, merely thinking that we are strongly conscious does not give us a reason for the claim that we are, in fact, strongly conscious. After all, Rey claims,

> ...the COG computer could be programmed to think that it, too, is strongly conscious -- even to think that it *knows for certain,* and incorrigibly, that it

> has a COG-transcendent consciousness -- while still (for most of us) failing to do so. (Rey, 1995, 130)

In other words, having the thought, 'I am strongly conscious' is compatible with being only weakly conscious, so the thought alone cannot give us a reason for ascribing strong consciousness to ourselves. Something more is needed.

Second, there is some reason to think that introspection is much less reliable than generally believed, and the way in which it goes wrong is suggestive for Rey. He cites a number of experiments that are designed to show that subjects are,

> sensitive to, but entirely unaware of, such factors as cognitive dissonance, prior expectation, numbers of bystanders, pupillary dilation, positional and "halo" effects, and subliminal cues in problem solving and semantic disambiguation. Instead of noticing these factors, subjects frequently "introspect" material that can be independently shown to be irrelevant to the causation of their behavior. (Rey, 1995, 125)

To take just one example (similar to one figuring into Churchland's thinking about introspection), subjects are asked to choose between two virtually identical pairs of socks, and they consistently prefer the pair on the right, despite

'introspecting' and reporting on non-existent differences in quality and design. A positional effect is at work, but subjects not only fail to realize this, they claim to introspect confabulated reasons for preferring one object to another to explain their choices.

Now Rey's point is not just the familiar one that factors of which we are not conscious sometimes contribute to our behavior. Some theorists suggest that when we think we are introspecting the real causes of our actions, sometimes what we are actually doing is applying theory about the sorts of things that are supposed to motivate people in the situation in which we find ourselves. (Nisbett and Wilson, 1977; Gopnik, 1993) For example, factors such as quality of material and craftsmanship ought to underlie preference, not position. So instead of introspecting the inner workings of our minds, we sometimes apply generalizations about human behavior and report the results when we claim to be introspecting, and we do this without realizing it.

We get so good at applying generalizations while growing up that we do not notice that we are not introspecting anything directly in these instances but constructing something theoretically. 'We are in the position of "expert" chess players, medical diagnosticians, or water dowsers who often take themselves to be sensing very sophisticated information in their domains "directly"', (Rey, 1995, 126) though what is really going on inside our heads is a series of

inferential steps, mediated by a number of generalizations.

According to Rey, we do not have to go along with the details of all of this just yet to appreciate the following crucial point:

> I might introspectively know that p without introspectively knowing *that* I introspectively know that p. Claims to know something introspectively may well require empirical support beyond any introspective claims themselves. (Rey, 1995, 127)

What we have access to, if anything, is not the nature of the access itself. The idea is that one might think that one can introspect p, but what is really going on could be something along the lines of the theoretical projections characterized above.

This is meant to tie into the claim that introspective grounds cannot motivate belief that we have strong consciousness, because, though one might think that one introspects strong conscious properties, introspection alone cannot tell us that we are, in fact, gaining access to those properties via introspection itself. For all introspection tells us, something else might be going on: the strong conscious properties that we introspect might be nothing more than theoretical projections, perhaps analogous to the projections going on when the positional effect underlies

choice. So the mere appeal to introspection does not on its own secure the claim that we are strongly conscious. We need reasons to believe that what we think is introspection actually is introspection.

Rey's third and perhaps most powerful reason for thinking that introspection cannot give us grounds for ascribing strong conscious properties to ourselves is that claiming to have direct experience of such properties is question-begging. What is needed is evidence that does not presuppose the existence of the thing in question. He writes:

> ...just as it's not enough in reply to an atheist to beat one's breast and claim that God exists because one has had direct experiences of Her, so it is not a reply to the eliminativist about strong conscious states to claim that one has direct, unquestionable experience of *them*....that is precisely what the atheist and the eliminativist are challenging. What is needed is evidence whose description does not *presuppose* the existence of the phenomena in dispute, but which nevertheless could not be *explained* without those phenomena. (Rey, 1995, 131)

When, for example, we claim that we know we are strongly consciousness because we have direct experience of strong conscious properties, we are assuming just what is at issue,

namely whether or not we are entitled to ascribe strong conscious properties to ourselves. Non-question-begging evidence is required if we are to have good reason for thinking that we are strongly conscious, but introspective reports presuppose the existence of just the properties in question. Therefore, introspection cannot serve as evidence for strong conscious properties.

The projectivist account

Rey calls the picture that emerges from all of this 'the projectivist account of consciousness'. He claims that problems which are of little significance in other subfields of philosophy take on special importance when considered as a part of the philosophy of mind. The issue of identity in connection with Theseus's ship is of little consequence to us when compared to the problem of personal identity, for example. In this instance, Rey maintains,

> ...we project an enduring object that corresponds (in our own case) to our personal concerns and (in the case of others) to the (more or less) standing effects they have upon us. (Rey, 1995, 137)

We do this despite the fact that there just is nothing in the world that corresponds to our projections.

Something similar is going on with respect to strong

consciousness, or so Rey maintains. He invokes a file-model of concepts: for each concept a person has there is something like a mental file with slots for information peculiar to the concept. For example, the mental file for the concept *Cat* contains information on the properties of standard cats and the conditions which usually obtain if the concept is properly employed. So too, Rey continues, with the concept *Pain*. The corresponding file contains information on the standard causes and effects of pain, recollections of past painful episodes, the standard behaviour of those who are in pain, and so on. The strong conception of pain arises, Rey claims,

> ...when we naively *think* about what are in fact the weak states we enter when the...pain concept is triggered....Just as we postulate a simple property of redness corresponding to the stability in our experience of red things, so do we postulate strong mental phenomena as the stuff of our immediate experience and as underlying typical behavioural effects in others. (Rey, 1995, 139)

Further, just as expert chess players make inferences so quickly and automatically that it seems to them as though they acquire information directly, it seems to us that we have direct or immediate access to strong conscious properties via introspection, when in fact we are projecting properties

beyond what is given. '*[W]e automatically deploy a concept that happens to involve commitments that exceed anything that introspection alone could possibly establish.*' (Rey, 1995, 139)

Even though the physical and computational facts about the brain do not warrant the postulation of COG-transcendent properties, we project them anyway. Strong conscious properties have far reaching implications for our moral life, among other things. So we are unwilling to take the scientists' word for it, as we might when told that there really is no single property in the world underlying red, and admit that there really are no strong conscious properties.

Rey's overall position might be understood as follows. Computational, functional and informational facts alone do not warrant the ascription of strong conscious properties, as our intuitions about the COG computer suggest. Further, we seem unable to give a micro-theoretical account of physical phenomena that upwardly necessitates strong conscious properties. The familiar strategy here is for the functionalist or physicalist to attempt to accommodate strong properties into his theory, perhaps by denying our intuitions about systems such as the COG computer (Lycan, 1995), or possibly invoking a distinction between kinds of knowledge that preserves the physicalist's claim that all facts are physical facts (Lewis, 1983).

However, Rey's strategy is to deny the existence of

strong conscious properties in the first place. He does so, as we have seen, by first arguing that there is no non-tendentious evidence for strong consciousness and then building up the projectivist account of consciousness in order to explain why strong properties are initially posited. This move effectively dissolves the problem of the explanatory gap by eliminating one of the things to be explained.

As we mentioned at the outset, there are two sets of intuitions in the literature about weakly conscious creatures, and considering them in more detail will help put Rey's position in focus. First of all, some maintain that if a zombie really is functionally identical to you, and if you are enjoying qualitative states, the zombie is as well. Some defenders of functionalism, in other words, deny the logical possibility of phenomenal zombies, creatures without qualia whose functional description is identical to that of a strongly conscious person. On the other hand, others take the possibility of a twin lacking qualia who is functionally identical to someone who has qualitative states as a counter example to the functionalist thesis that the nature of mental states is exhausted by functional descriptions. If the two creatures are functionally identical but mentally different, then functionalism does not capture the nature of mentality.

Now given the projectivist account of consciousness and what Rey says about the COG computer, Rey does not fit precisely into either of these camps. He agrees with the

anti-functionalists, in so far as they claim that zombies are possible, but denies that this undermines functionalism. Instead of arguing that a zombie who is functionally identical to you shares your qualia, he maintains you actually have no qualia. There really are no qualia to be had.

His reason for this claim is based on the available evidence for consciousness. The argument might be more clear if cast in terms of perspective. Facts available from the third-person point of view -- facts about the nature of the brain, computational facts, behavioural facts and so on -- support only the ascription of weak consciousness. Rey's characterization of the explanatory gap and the intuitions he invokes about the COG computer, in particular, our reluctance to ascribe strong consciousness to it, makes this plain enough. First-person facts -- facts available to us via introspection -- are inadequate as well. If first-person and third-person facts are all we have, and if neither one gives us evidence for strong consciousness, then there just is no clear evidence for the existence of qualitative properties.

2.4 The Plausibility of the Projectivist View

Does Rey's argument force us to accept this conclusion? Let us grant that there is an explanatory gap, that physical facts alone do not warrant the ascription of strong conscious properties, provided we bear in mind that we are talking about the physical facts as we know them. It may be that a better

conception of the physical facts, or indeed, more physical facts, would issue in an understanding of the mental in terms of upward causation, determination, entailment, or something not yet conceived. At any rate, dealing with the facts as we know them is the best we can do, and given what we currently believe, we cannot give an account of consciousness in the way that we can give an account of the expansion of water when it freezes. There is an explanatory gap.

Now it might be claimed that while there is nothing about observing brains alone that leads us to believe that there is strong consciousness, some third-person facts, behavioural facts, really do give us a reason for positing the existence of qualitative states. For example, when a person with a tooth ache groans, it seems reasonable to suppose that he or she is experiencing pain qualia. The problem with this is that, if it is true that zombie replicas are behaviourally indistinguishable from twins who have qualia, then reactions appropriate to experiencing qualia are not good enough to establish the existence of qualia. A psychological conception of mind, not a phenomenal one, is all that the physical evidence will bear.

The interesting part of the argument, for our purposes, is the second half, the claim that introspection does not warrant the ascription of strong conscious properties in our own case. Rey's question, as we have seen, is: what grounds do we have for the self-ascription of strong consciousness?

What we want to say -- and as we have seen what many philosophers do say -- is that we have direct access to strong properties via introspection, and that is how we know that we are strongly conscious. Let us have another look at Rey's three arguments against this move, with a view to determining whether or not his conclusion really is warranted.

Response to Rey's three claims about introspective evidence

His first claim is that merely having the thought that we are strongly conscious does not give us grounds for thinking that we are, in fact, strongly conscious. A weakly conscious computer could be programmed to have a thought with that content too, so the thought alone is not proof that we are strongly conscious. If you have difficulties believing that a computer might have a thought, in particular the thought that it is strongly conscious, it might help to imagine a zombie thinking that it is strongly conscious. If it really could be in the same functional state as you but fail to have qualia, then it would have the same thought that you do when you think that you are strongly conscious. It seems that just having the thought 'I am strongly conscious' does not secure the claim that one is, in fact, strongly conscious.

However, it is not clear that anyone who points to introspective access as evidence for strong consciousness is pointing to thoughts about strong conscious properties, particularly the thought, 'I am strongly conscious'. Instead,

as we have seen, it is claimed one knows one is conscious because introspection gives us direct access to strong properties themselves. The evidence is not thoughts about the access, nor is it thoughts about what it is claimed we have access to -- the evidence for qualitative properties is what the access puts us in touch with in the first place. Rey's claim seems to focus on our thoughts about strong consciousness, but that is not the source of the qualiaphile's evidence.

The person who argues that he knows he is strongly conscious is not claiming that he knows this simply because he has the thought that he is strongly conscious, but because, for example, some things seem green to him. Maybe Rey is right and there really are no strong consciousness properties, only projections. Whether or not this is so, it is on the basis of the awareness of green things, whatever that turns out to be, that some have the thought that they are strongly conscious. It is not the thought that is the basis of the qualiaphile's insistence, but the strong properties he claims to detect directly. So claims about the thought that one is strongly conscious might miss the point.

Of course, Rey could respond by arguing that the alleged direct experience of qualia is question-begging, but this takes us away from the first argument and into the third. We will consider it in a moment.

Rey's second argument against the proposition that we

introspect strong conscious properties turns on his claim that 'I introspect that p' does not iterate, that is, 'I might introspectively know that p without introspectively knowing *that* I introspectively know that p'. (Rey, 1995, 127) Rey's skepticism at the moment is, once again, not aimed directly at what we are introspecting when we claim to be introspecting strong conscious properties, but what we are doing when we claim we are introspecting, when we claim to be detecting strong properties directly.

Rey is first undermining our confidence in the access, claiming more or less that what we think we are doing when we introspect is not really accessing anything at all. If he is right and we are really not introspecting when we claim to be introspecting, then we are really not detecting strong conscious properties directly. If we are not doing that, then our claim to have evidence for strong consciousness via direct access to it is clearly undermined.

What else could we be doing when we claim to be introspecting? His remarks about the sometimes theoretical basis of introspective reports postulated by researchers, coupled with his own projectivist account, provides an alternative. When, for example, the positional effect leads a subject to say that she prefers one more or less indistinguishable tin of soup to another, but her 'introspected reason' is that it is higher quality soup, what is really going on is not introspection but theoretical

projection. Rey's point is that introspection alone does not tell us whether it is introspection or projection that is going on, so introspective claims alone cannot give us grounds for thinking that we introspect strong conscious properties. Further evidence is required to prove that what we take to be introspective reports are actually the result of detecting strong conscious properties, not projection.

Before the projectivist possibility came into view, the fact, if it is a fact, that 'I introspect that p' does not iterate was not problematic. If it were asked, 'how do you know that what you are doing is introspecting', it was reasonable enough to reply, 'what else could I be doing?' Skepticism about introspection only sinks in after there is another possibility. We need to ask, then, whether or not it is a live possibility, whether or not it is possible that, each time we think we are introspecting, projection might be going on.

Searle makes some remarks in a similar context that are illuminating:

> ...we do not *postulate* beliefs and desires to account for anything. We simply experience conscious beliefs and desires. Think about real-life examples. It is a hot day and you are driving a pickup truck in the desert outside of Phoenix...you want a cold beer so bad you could

> scream. Now where is the "postulation" of a desire? Conscious desires are experienced. They are no more postulated than conscious pains. (Searle, 1992, 59)

It might be that some of our introspective reports are really projections, but Searle reminds us to think of everyday cases. Is it really possible, as it has to be on Rey's view, that every time we introspect a projection could be going on -- worse, that every time we think something seems green or our thirstiness seems to us overpowering, we are not introspecting, but theorizing about how people are supposed to think about thirst and projecting the results on our own conscious states? Is it a real possibility that the pain one feels after banging a knee on a desk is really a projection? Is it a real possibility that *every* pain ever felt is really a projection?

What we may be doing here is beating our breast, like the theist in Rey's third counter argument, who claims that he just knows that god exists because he has direct experience of her. What is needed, Rey claims, is non-question-begging evidence for the existence of strong conscious properties, evidence that does not presuppose the existence of the thing in dispute. What would Rey count as non-tendentious evidence for strong properties?

If the experience of strong conscious properties is tendentious, it might be worth wondering whether Rey's

criterion for acceptable evidence is too strong. Is sensory evidence of the external world, for example, acceptable evidence for the view that the external world exists? It seems to be evidence that presupposes the existence of the thing in question. One might argue that the claim, 'I know that the external world exists because I see it', ought to be as problematic for Rey as the claim 'I know strong conscious properties exist because I introspect them'. Or it might be claimed that using sensory evidence as grounds for belief in the external world presupposes the existence of eyes and ears, reflective surfaces and sound waves -- surely all of this begs the question, on Rey's view. Are we not led to the conclusion that, if Rey's conception of evidence is what we are working with, we have no evidence for belief in the external world, including the existence of Rey, the COG computer, and the argument against strong conscious properties?

He gives us a reply to this sort of worry and an answer to the question, what counts as non-tendentious evidence, in a footnote:

> There are non-question-begging reasons to believe in the existence of material objects -- they afford the best account of our persistent thought that there are [material objects] -- that are lacking in the case of COG-transcendent consciousness. (Rey, 1995, 132)

Rey is saying that an inference to the best explanation counts as a non-question-begging reason for belief in material objects, but that such an inference is not motivated with respect to strong consciousness. Is this true?

We have two possible explanations for what is going on when we claim to introspect strong consciousness: introspection of real properties or the projection of nonexistent properties. Which of these is the best explanation for our persistent thought that, for example, some things seem green?

If it turns out that the best explanation for what is going on when we introspect is direct access to real properties and not projection, then we have a response to both Rey's second and third arguments. The second argument generates skepticism about strong consciousness only if it really is possible that all of our experiences of qualitative properties are actually projections. However, if it is reasonable to suppose that the best explanation for what is going on when we introspect is direct access to strong properties, then it is reasonable to suppose that strong properties exist, even if Rey is right and introspective reports do not iterate. The third argument generates skepticism only if there is no non-question-begging evidence for strong consciousness.

What is needed is a good reason, a non-question-begging reason, for thinking that introspective access really is

direct access to strong properties. If the best explanation for what is going on when we introspect is direct access to strong properties, not projection, then we have a non-question-begging reason for believing in strong properties, by Rey's own criterion. If we have that, we would have a reason to reject the claim that all of our experiences of qualitative properties are really projections. We would also have a reason to deny the claim that there is no non-tendentious evidence for strong consciousness. We would have a reason for believing that strong consciousness exists.

What best explains our belief in the existence of strong properties? What best explains our persistent thoughts about the part of the explanatory gap that Rey wants to eliminate: 'something's looking green...'the way things feel'...'qualia''? (Rey, 1995, 128) Why, in short, do we think that things seem a certain way to us?

On the one hand, the view is that our persistent thought that some things seem the way they do is, in every case, a theoretical projection. Pains do not really hurt and things do not really seem green. What is going on, in every case, is an inference from certain generalizations that we have learned as we mature about how people are supposed to behave when they are in pain or when they see green. Crying babies, on this view, do not cry because pain hurts -- presumably, they have not yet learned the necessary concepts with which to project the nonexistent hurtfulness of pain onto their sensory states.

What is there to recommend this view and override what seems the counterintuitive nature of the position?

This view affords some measure of epistemic relief from the explanatory gap, but other questions seem to outweigh that gain: if there are no strong conscious properties, why should we ever have developed a theory about people that includes such things in the first place? Rey suggests that we might have a hardwired proclivity for thinking of ourselves and others as strongly conscious, but what possible evolutionary explanation could one give for such a suggestion?

Expert medical diagnosticians and chess players spend a considerable amount of time learning the generalizations that drive the inferences they make; so when do we learn that pain hurts? Given the great difficulty we sometimes have describing conscious properties, how could one even begin to teach someone else what it's like to see green? The epistemic relief from the problem of the explanatory gap that Rey's dissolution affords seems to vanish quickly once we start thinking about the details of his projectivist proposal.

According to the more intuitively plausible view, we think that some things seem green to us because they really do. The view accords well with prior practice, and it preserves the truth of everyday explanations and descriptions of our behaviour that invoke qualitative properties, and indeed a large part of our general conception of what it is to be human -- all of which has to be rewritten on the

projectivist view. On this view, to take a very mundane example, it makes sense to say that we groan because the toothache hurts, but that explanation does not work according to the projectivist. Though taking up this view leaves us with an explanatory gap, it does not saddle us with the number of unanswered questions the other view does. The gap, on this view, does not exist because we posit nonexistent properties, but because, like so much else, there is something about brains that we do not yet understand. It is a gap we can cope with, not grounds for eliminativism.

In the end, we have a response to Rey's argument. Given what counts, on his view, as non-question-begging evidence, we have evidence for the claim that we have strong conscious properties. The everyday understanding of ourselves, the understanding that builds strong conscious properties into our conception of ourselves and others, is the best explanation for our persistent thoughts that we have strong conscious properties. If such an inference is motivated with respect to the existence of material objects for Rey, it seems likely that he will have to accept a similar inference with respect to strong conscious properties too. By his own account of the nature of good evidence, there is evidence for strong conscious properties, properties we can still hope to understand.

3. Is Consciousness Important to Theorizing About the Mind?

Thus far, we have considered two arguments for the elimination of consciousness: Churchland's argument that a certain conception of consciousness does not refer and Rey's argument for the claim that there is no evidence for strong conscious properties. There is at least one more position in this neighbourhood that is worth considering -- if not really in this neighbourhood, then perhaps somewhere midway between eliminativist views and the realist but sceptical views we will consider in the last three chapters. It is possible to take up a realist stance with respect to consciousness, agreeing that the everyday concept refers, but nevertheless maintain that consciousness is unimportant to theorizing about the mind. It may be that we are 'getting at' something when we talk about conscious states, but the questions that concern us here, what is consciousness and how is it related to the world, are red-herrings from the point of view of serious empirical and scientific work.

A good representative of this view is Wilkes (see Wilkes 1984, 1988, 1995), who maintains that there really is no great problem of understanding consciousness, 'that in fact consciousness as such is not at all important, and that psychology and the neurosciences would lose nothing, and gain much, by refusing to chase this will-o'-the-wisp'. (Wilkes, 1984, 224) Though her interest is primarily with the application of the notion of consciousness in the behavioural

and brain sciences, she maintains in the end that 'common sense psychology and the philosophy of mind need not bother with the notion either'. (Wilkes, 1984, 224) Wilkes claims that a clear general conception of consciousness that might inform scientific and philosophical theorizing about mental phenomena cannot be had. Therefore trying to understand the mental (in particular states like being awake, sensations, perceptions, and propositional attitudes) by using the notion of consciousness is wrongheaded.

Her argument for this claim begins in a discussion of what she takes to be four different uses of the word 'conscious' in the vernacular, and we will consider this discussion in detail in part one of this chapter. She makes a very large number of claims at the outset, piecing it all together as she goes along, and there is no tidy way to recapitulate all of this and interpret it at the same time. Therefore I plan to limit myself primarily to recapitulation in the first part of this chapter.

It is on the basis of her discussion of the vernacular uses of 'conscious' that Wilkes concludes that consciousness is 'a *prima facie* unpromising phenomenon for systematic exploration.' (Wilkes, 1984, 229) In part two, I will try to explain the connection between her description of everyday uses of 'conscious' and this conclusion. This will necessitate some consideration of the differences Wilkes identifies between common sense psychology and scientific

psychology, and what the word 'conscious' does and can be expected to do in each. Her view is that 'conscious' as we use it in our everyday lives suffers from certain defects that make its referent an extremely unlikely candidate for scientific study.

These defects will also concern us in the third part of this chapter. There I try to build a case against Wilkes's claims. In the end, I hope to show that Wilkes does not provide good reasons for believing that the concept of consciousness is unimportant to theorizing about the mind or everyday claims made about our mental lives, and her view that consciousness is not the sort of phenomenon that can be understood.

3.1 Four Uses of 'Conscious'

It is worth being prepared at the outset for the fact that Wilkes never says exactly what the problem of understanding consciousness is. This might be because she believes that there really is no problem of consciousness. It would be helpful, though, to have some idea of what the alleged pseudo-problem is, what others might mistakenly suppose is a real problem, according to her. Given the focus of her articles, she understands the problem of consciousness as the problem of understanding the relation between the brain and conscious states. Though she claims that a number of different states are sometimes called 'conscious', she focuses

on bodily sensations, sensory perceptions, and propositional attitudes and the general state of wakefulness. Presumably, the problem of consciousness is understanding the relation between such states and the brain.

Her understanding of 'consciousness' and its cognates is also never set forth explicitly in the text. Instead we have a discussion of different sorts of things that we sometimes call 'conscious' and little straightforward characterization of what 'conscious' itself actually means. However, she provides two remarks about what consciousness is not.

First, she says, it is misleading to think of consciousness as,

> ...a kind of internal illumination; a spotlight in the private theatre of the mind that picks out some, but not all, of the passing show; a light that may be dimmed or intense, loosely or narrowly focused, but which is unambiguously either on or off. (Wilkes, 1984, 229)

The metaphor is unhelpful, she argues, because it conflicts with the most common use of the term 'conscious'. Most of the time, when we claim to be conscious of something, we are talking about our consciousness of things in the world and not making reference to throwing the spotlight of consciousness on some inner state. The inner illumination metaphor does seem to

fit certain rare and sophisticated sorts of attending to one's experiences (introspection, perhaps), Wilkes says, but most of the time when we talk of being conscious of something, we are not engaged in this sort of attending. Most instances of being conscious of something, therefore, fail to fit the model.

The second unhelpful intuition about consciousness is 'the idea that consciousness is some form of attention.' (Wilkes, 1984, 230) Her objection to this view is based on Dennett's claim that 'any problem-solving or game-playing computer pays attention' (Dennett, 1978, 209) despite not being conscious. Wilkes does not offer a defence of this view, but concludes from it that, if some attending goes on in the absence of consciousness, we cannot understand consciousness by thinking of it as a kind of attention.

These two negative claims are all Wilkes offers in the region of a characterization of consciousness, and no doubt this reflects her sceptical view of consciousness and its importance.

A certain pattern emerges in her discussion of different uses of 'conscious'. Wilkes first characterizes the thing thought to be conscious (people who are awake, sensations, perceptions, and attitudes), and then advances a variety of claims about the utility of the concept of consciousness for our understanding of the things in question. We are provided with no account of the sense in which the alleged four uses

are different uses. We will consider some of the difficulties that arise as a result of this in a discussion of heterogeneity in the third part of this chapter. For now, though, let us put this worry aside and consider Wilkes's treatment of the four things called 'conscious'.

Being awake

In what might be its most straightforward use, she says, 'conscious' is a one place predicate applied to complete systems -- people, organisms, and maybe robots, according to her. To use her example, it might be said that 'John is conscious now'. The other three uses of 'conscious' are either relational (for example, 'John is conscious of Betty's footwear') or one place predicates of mental events (for example, 'a conscious desire'). According to Wilkes, linguistic convention has it that 'conscious' used in the first sense means being awake, as opposed to asleep or comatose.

However, Wilkes claims that there are a great number of states involved in the sleeping/waking cycle that the conscious/non-conscious dichotomy does not accommodate. She argues that sleeping divides into several kinds of states, some of which are more like states of wakefulness than those most of us call states of sleep. In the case of dream states, for example, she says that 'the conscious/non-conscious dichotomy breaks down completely....' (Wilkes, 1984, 230)

There are grounds for calling dreamers conscious, she says, since dreamers have thoughts and other experiences associated with conscious states; dreamers even have something very much like the experience of perceiving things around them, she claims, though the sense in which 'experience' is used here is not made plain. Yet there are grounds for calling dreamers unconscious too, since they are normally largely unaware of their environment.

Sleep walking and talking present similar problems for the dichotomy, she maintains, because both occur during states of deep sleep but involve what would otherwise be characterized as sophisticated conscious behaviour: using language and negotiating the environment.

More exotic phenomena do not fit into the dichotomy easily either, according to Wilkes. For example, epileptic automatism is a condition brought on by a seizure which seems to leave the subject in a sleep-like state. However, if the subject is performing some well-rehearsed or mundane task at the time of the episode, like driving a car or playing the piano, she may continue with the activity as though sleepwalking through it. The many layers of hypnotic states might be even more difficult to classify, Wilkes argues, because each one seems to share so much with both conscious and non-conscious states.

The difficulties associated with using the conscious/non-conscious dichotomy in making sense of ordinary sleeping

states and exotic phenomena like automatism and hypnosis lead Wilkes to conclude that,

> ...classification [of these states] in terms of consciousness or its absence is simply too crude to cope with all this diversity; surely profitable research into these and other phenomena will require a theoretical classification determined along different, and probably more fine-grained, principles. (Wilkes, 1984, 231)

The point is, it seems, that 'conscious' is too imprecise to help make the kind of distinctions needed for research to get a foothold.

The obvious question is, what prevents greater precision being built into the notion of consciousness as a result of further research and our own linguistic decisions? Some sleeping states may seem much like states of wakefulness, given our current understanding, just as some mammals seem much like fish to a person who is not well versed in biology. What prevents us from coming to a more precise conception of consciousness, as we have with the concept of mammal, that will enable us to sort through the diversity Wilkes rightly identifies? We will take up this point in the final section of this chapter.

Sensations

The second use occurs when 'conscious' modifies bodily sensations. Her discussion of sensations begins with a recapitulation of a traditional philosophical characterization, rooted in familiar epistemological criteria. Sensations are allegedly private, which is to say that the same sensation cannot be shared by two or more individuals. It is also said that a person in pain knows about it immediately -- he need not follow a chain of inferences to know that he is in pain. It is in virtue of this unmediated access to sensations that some philosophers maintain that our claims about our own sensations are incorrigible, or so Wilkes maintains.

Because each of these aspects is the subject of ongoing dispute, it might be best, she says, to identify sensations ostensively: they are things such as pains, itches, tingles, butterflies in the stomach, pins and needles, tickles and twinges. She claims that an examination of the members of this list reveals at least four shared features that might be used to further characterize sensations.

All sensations have temporal duration. Butterflies in one's stomach might begin shortly before a lecture and end once it is delivered. Sensations also have felt locations -- for example, one might feel a stabbing pain in the lower right side of the back. Further, sensations can typically be gauged with reference to a qualitative scale. A pain might be

experienced as faint rather than intense. Finally, she notes that, according to our linguistic conventions, nothing seems to count as a sensation unless we are conscious of it. For example, nothing counts as a pain unless it feels a certain way to the subject. For sensations at least, *esse est percipi*. It is this feature of sensations that Wilkes focuses on the most.

She says that the most promising line of research into sensations treats pain and the rest in terms of causal or functional roles. She asserts that the physical occupants of functional roles cannot be identified as single states, at least some brain research suggests, but as sequences of states throughout the brain. Whatever it is that occupies the role of pain in the human brain, for example, it is not just the firing of C-fibres, but a series of information processing states ranging through many different sub-systems in the brain.

A functional characterization of pain takes into account not just these states, but the causes and effects of such states. In humans, Wilkes observes, the effects of pain very often involve highly complex behavioural responses, in conjunction with the activation of higher cognitive functions and the formation and behavioural expression of the characteristic beliefs and desires that accompany tissue damage.

However, Wilkes notes, sometimes we react immediately to

tissue damage, withdrawing our fingers from danger 'without thinking' -- before our higher cognitive functions have the chance to start forming beliefs and desires. In other cases, higher cognitive functions are never activated at all, as in the case of a person's unnoticed positional shifts while sitting in an uncomfortable chair. This suggests to Wilkes that the initial stages of sequences of states throughout the brain that occupy the functional role of pain can be set in motion and issue in behaviour before the requisite information reaches the parts of the brain that subserve higher cognitive function, or even in the absence of information affecting higher cognitive functioning at all.

So if the most promising line of research into sensations really is in terms of functional role, and if the multiple occupant story Wilkes is endorsing here is correct, then it might make sense to recognize a class of mental events which she calls 'unnoticed pains'. Such pains, if they are pains, are caused by tissue damage and issue in many of the same physical events as standard pains do, save for the activation of higher cognitive functions and the behaviour that attends such activation. This might be enough for it to make sense, from the point of view of research, to identify unfelt pains as pains.

At least part of Wilkes's point here is that profitable lines of research in the brain and behavioural sciences might not take into account the vernacular conception of conscious

sensations and the linguistic convention that, for sensations, to be is to be perceived. She writes,

> I am suggesting that examination of the functional description of pain may, and probably will, require that these instances of unnoticed pain count, properly and equally, as pain discrimination by the organism despite the fact that we need not be aware either of the painful stimulus or of the compensatory response; and hence that the sciences may for principled reasons refuse to agree with the vernacular that the *esse* of pain is *percipi*. (Wilkes, 1984, 233)

Why would that matter? The point is, for Wilkes, that the way linguistic convention has it at present is that unconscious or unnoticed sensations *cannot* exist. This could be a mistake, and a fundamental one, according to Wilkes. It looks like there are very good reasons for researching into the nature of unnoticed sensations, so 'conscious' as it applies to sensations not only fails to help research, it 'obfuscates and confuses one's ability to describe and explain other sorts of pain.' (Wilkes, 1984, 233)

She also considers the dissociated pain characteristic of some cases of hypnotic anaesthesia. A suitably prepared hypnotized subject is told that he will feel no pain and then

has a hand placed in freezing water, an experience reported as extremely painful by the unhypnotized. The subject sincerely verbally reports that he feels no pain, but, if given a pen and paper, some subjects will offer a simultaneous written commentary about the intensity of the pain.

Following this discussion of both dissociated pains and unnoticed pains, Wilkes offers a general conclusion about bodily sensations:

> Such forms of dissociation, common enough in hypnosis and elsewhere, make nonsense of the attempt to describe what is going on in terms of consciousness or its absence. Nor seems there to be a fact of the matter about whether the hypnotized subject is *really* in pain or not. After all, the vernacular concept of pain is a pretty crude notion which includes a heterogeneous diversity of sensations (few sensations are less similar than a dull stomach ache and the prick of a pin); it is only to be expected that a scientific taxonomy should delineate the extension rather less crudely and -- perhaps -- in accordance with theoretical demands of its own which may include unnoticed and dissociated pain as legitimate kinds. (Wilkes, 1984, 234)

We have, then, a reiteration of the conclusion reached with respect to 'conscious' understood as 'being awake': the concept of consciousness is too crude to cope with diverse mental phenomena. There are two new claims here as well. First, thinking in terms of consciousness or its absence interferes with our ability to describe and explain non-standard cases, like dissociated and unnoticed pains. Second, Wilkes stresses here the heterogeneity of phenomena lumped together by the vernacular concept. Similar conclusions apply to sensory experience, or so Wilkes suggests.

Sensory experience

Wilkes notes that 'conscious' is applied to a third group of things, sensory experiences: seeing the moon, smelling a cow, and so on. She explains why this is a different usage of 'conscious' by arguing that sense experience does not share the four features characteristic of sensations.

She argues that sense experience is typically continuous, not disjointed as something like a sudden pain might be said to be. It also has no location, unlike butterflies in the stomach. According to Wilkes, one cannot place sensory experience on a qualitative scale as one might with pain -- that is to say that visual experience, for example, cannot be faint or intense like sensations sometimes are. Further, though the vernacular has it that the existence of sensations consists in their perception, she claims, the same cannot be

said of the everyday understanding of sensory experiences.

Given phenomena such as subliminal perception and blindsight, it is wrong to claim that sensory experience must be perceived in order to exist. Exactly why she believes phenomena such as blindsight and subliminal perception but not hypnosis and automatism have altered the common sense use of the term is not made clear, and we will return to this point in due course.

So only some sense experience is conscious. The difference between conscious and un- or non-conscious perception is difficult to delineate, according to Wilkes. Using the ability to report as a basis for the distinction, for example, is dubious because it operates with a built in bias against creatures and sub-systems without the ability to use language. This seems like a mistake, she argues, because it is likely that creatures without the ability to report have some share in consciousness -- so too, perhaps, with sub-systems in the brain. More importantly, she claims, even amongst language users, reporting leaves us with ambiguous results because asking someone to report inner states alters the phenomenon under consideration.

Much like the case of sensation, the best we can do in our efforts to characterize conscious and unconscious sensory experience, she maintains, is invoke examples of a few clear cases of each. Despite her worries about reporting, we are directed to the reports of a subject carefully attending to

her visual field for an example of conscious experience and to cases of blindsight and subliminal perception for examples of non- or unconscious experience.

Why blindsight and subliminal perception count as examples of experience is never made clear. As we saw in our consideration of Churchland, it is prudent to be wary of philosophical interpretations of psychopathological cases, and it is not yet clear how best to characterize instances of subliminal perception. Whether or not such things should be taken as instances of non-or unconscious experiences, and whether or not the very notion of a non- or unconscious experience makes sense, is a difficult question, and we will return to it in chapter five. However, we get little in the way of explanation of Wilkes's line of thinking here.

She goes on to claim that using paradigm cases to mark the boundary between conscious and non-conscious sensory experience suffers from a certain sort of defect:

> To resort to this way of marking the distinction is, however, to admit that the bulk of our sensory experience (including the most common and interesting forms), and practically all the sensory experience of other animals, elude or resist the dichotomy. (Wilkes, 1984, 227)

She claims that a person driving a car 'on automatic pilot'

while carrying on a conversation leaves one wondering whether or not his experience of the road is conscious or unconscious in some sense. Much of our everyday experience is like this, she says, and concludes that our current use of 'conscious' leaves a great deal of sense experience neither clearly conscious or non-conscious.

Propositional attitudes

Finally, Wilkes says, 'conscious' is applied to a fourth group of things which, much like sensations, is itself heterogeneous: the propositional attitudes. By propositional attitudes, it is clear that she means the usual set of things like beliefs, desires, hopes and so on. Her discussion focuses on two distinctions that might be made among the attitudes: dispositional versus occurrent attitudes and self-conscious versus non-self-conscious attitudes.

The first distinction is partly clarified with examples. Occurrent attitudes are evidently in existence when one engages in explicit deliberation or is suddenly struck by some thought or other. Dispositional attitudes, she says, 'are the beliefs, desires, memories, aversions, preferences (*etc.*) which we can ascribe fairly to the sleeping man, or to an animal.' (Wilkes, 1984, 228) Here, Wilkes is talking about propositional attitudes that might be ascribed to a person, though that person is not currently explicitly entertaining them.

Wilkes suggests that the belief or desire in question might just follow from other beliefs and desires that the person is known to hold, whether or not the person has made the necessary inferences to arrive at the dispositional belief or desire ascribed to him. We might ascribe the dispositional belief 'Clinton was born in America' to a person who believes both that Clinton was born in Arkansas and that Arkansas is in America -- even though the person is not now entertaining (and perhaps never has entertained) the belief that Clinton was born in America.

The second distinction is between self-conscious attitudes and non-self-conscious attitudes: 'those that take, and those that do not take, the agent himself or his own mental states as their subject-matter.' (Wilkes, 1984, 228) Examples of the former include my belief that I am sleepy; examples of the latter include my hope that there is still some coffee in the kitchen.

Her interest here is whether or not these two distinctions might somehow be used to understand the difference between conscious and unconscious propositional attitudes, and her conclusion is that the distinctions are no help. Numerous examples are marshalled to show that, for example, an occurrent belief might be either conscious (when one engages in explicit deliberation) or unconscious (when, say, a person who has just fallen down some stairs claims to have thought that there was a another step). Though the

latter looks very much like a dispositional belief, the claim is undefended.

She also argues that one cannot distinguish between conscious and unconscious beliefs by claiming that the former are self-conscious and the latter are not. A person can have beliefs about herself that are not necessarily conscious -- for example, the anorexic who seems to have the non-conscious that belief her body is fatter than it really is. A person can also have beliefs that are not about herself that are conscious -- for example, my belief that this glass is empty.

If these two distinctions cannot help us distinguish between conscious and unconscious attitudes, then, Wilkes argues, much like the other ascriptions of 'conscious', all we have to go on are a few paradigm cases. Thinking in words, explicit deliberation and the like are examples of conscious beliefs, and the desires detected by Freudian psychoanalysis are examples of unconscious desires. We will consider the speed with which this conclusion is reached in part three of this chapter.

Parallelling her earlier claims that much mental phenomena is not clearly conscious and not clearly unconscious, Wilkes maintains that 'the vast majority [of attitudes] evade the dichotomy.' (Wilkes, 1984, 229) A few paradigm cases are the best we can do.

Why? Wilkes claims that,

(in our common sense psychological descriptions and explanations) we are inconsistent and vague about the applicability of the adjective 'conscious' to the majority of the propositional attitudes. It is thus *prima facie* implausible to suppose that a blurred, shifting conscious/non-conscious dichotomy which is neither exhaustive nor exclusive could hold much promise for systematic study. (Wilkes, 1984, 236)

Wilkes's conclusions

Wilkes draws a number of conclusions about the utility of using the notion of consciousness in our efforts to understand the four phenomena under consideration. It will be useful to have a systematic expression of her conclusions if we are to come to terms with her argument.

1. *Being awake*: 'Classification in terms of consciousness or its absence is simply too crude to cope' (Wilkes, 1984, 231) with the different stages in the sleeping/waking cycle and other, related phenomena. According to Wilkes, the common sense conception of consciousness is not fine-grained enough to enable us to classify confidently a number of different states as either wakeful or part of the sleep cycle.

2. *Sensations*: The vernacular concept of consciousness as applied to sensations is 'a pretty crude notion'(Wilkes,

1984, 234), which obscures research into potentially profitable areas, like the study of dissociated and unfelt pain. Further, the general class, conscious sensations, names a hodge-podge. What could be more different, Wilkes asks, than the prick of a pin and a dull stomach ache?

3. *Sensory Experience or Perceptions*: All we have are a few paradigm cases to help us distinguish between conscious and unconscious perceptions. According to Wilkes, this leaves 'the bulk of our sensory experience' (Wilkes, 1984, 227) uncharacterized in terms of the presence or absence of consciousness.

4. *Propositional Attitudes*: Again, because we must have recourse to paradigm cases, 'the vast majority' (Wilkes, 1984, 229) of the attitudes are neither clearly conscious nor clearly unconscious. Further, the class of propositional attitudes is (like sensations) a hodge-podge.

In a word, the general view is that 'consciousness' names a heterogeneous set of things, and it does so with little precision.

Immediately following her consideration of the four uses of 'conscious', Wilkes arrives at the following general conclusion:

> From the preceding section, consciousness emerges not only as thoroughly heterogeneous, but also as a *prima facie* unpromising phenomenon for systematic

exploration. (Wilkes, 1984, 229)

Presumably, this is a variation on her general conclusion that consciousness is unimportant to theorizing about the mind: unpromising phenomena just are unimportant phenomena with which the sciences need not bother. It seems a very large jump from the conclusions reached with respect to the various things called 'conscious' to this general conclusion about the utility of the notion of consciousness.

Even if things are as bad as Wilkes maintains -- even if, for example, characterization in terms of 'conscious' or its absence leaves the bulk of the attitudes uncharacterized -- why think careful conceptual and empirical work cannot improve our understanding of consciousness? Why think that this project is a red-herring? The burden of section two is to interpret her general conclusion and determine how Wilkes makes what seems a rather large jump. In section three we will try to see whether or not it is warranted.

3.2 'Conscious' in Scientific and Common Sense Psychology

A clear understanding of the line of thinking Wilkes is pursuing here requires a consideration of Wilkes's distinction between common sense (or folk) psychology (CSP) and scientific psychology (SP), and her understanding of the role of natural kinds in each.

Common sense and scientific psychology

Wilkes's conceptions of CSP and SP emerge as she contrasts them. She does so in terms of the aims and language characteristic of each, as well as the place of laws and natural kind terms in each.

The fundamental aim or task of any science, including a science of human psychology, Wilkes argues, is the systematic description and explanation of suitable phenomena (we will return to the notion of suitable phenomena in a moment). CSP, on the other hand, sometimes engages in systematic description and explanation, but, she says,

> Unlike SP...the vocabulary of CSP needs to cope with tasks other than systematic description and explanation: joking, jesting, jeering, hinting, hassling, hustling, commending, condemning, consoling, blaming, bullying, threatening, reassuring, encouraging, sympathizing, proselytizing, punning, warning...and such an Austintatious list could continue long. (Wilkes, 1995, 101-2)

Because CSP and SP have different aims, the nature of the terms employed by each differ greatly. CSP is very good, she says at predicting, explaining and describing in very specific instances -- what this particular person is doing with respect

to that one, at a particular time and place. Its specificity consists in the loose nature of its language -- among other things what matters is the fact that CSP terms are highly context sensitive and shift meanings as the result of subtle nuances. She argues,

> [t]he specificity, exactness, accuracy of CSP come largely from *lack* of well-defined precision: from the way shades of nuance in the terms chosen, tone of voice, context...all contribute to every explanation or assertion. (Wilkes, 1995, 102)

Because its terms have no very precise meaning unless enmeshed in a certain context and expressed in a certain way, CSP terms differ fundamentally from the more precisely delineated SP terms. Because SP is in the business of systematic description and explanation only, its terms have to be context and speaker insensitive. For SP, the more clearly demarcated the terms and the more general their application, the better. Just the opposite holds true for the vocabulary of CSP, or so Wilkes maintains.

Another difference Wilkes identifies is that generalizations and laws are not very important in CSP, while they are crucial for SP. We acquire a working knowledge of CSP just in virtue of being competent speakers of a language -- learning explicit laws and generalizations is not a large

part of the practice of CSP. This is not to say that there are no regularities underwriting our use of CSP terms, but Wilkes's point is that we do not need to know what they are to engage in CSP. A certain kind of know-how is all that is required. However, SP just is the business of explanation and description in terms of explicit laws deliberately sought after and tested.

This distinction between the implicit and explicit use of laws partly underwrites Wilkes's claim that only certain phenomena, natural kinds, are suitable for the kind of systematic study characteristic of SP.

Natural kinds and interesting laws

What makes a given grouping a natural kind, on her view, is that members of the grouping are members just in virtue of the fact that they are subject to laws.

The following passages are instructive:

> SP needs to have or to seek clear and unambiguous *explananda*-phenomena; phenomena such that it makes sense to suppose that we could get laws about them. (Wilkes, 1995, 104)

> ...'natural kinds' [are] the joints into which Nature is carved. Natural kinds, in short, are systematically fruitful *explananda* and *explanantia*,

> where members of the kind are held together and governed by law(s).... (Wilkes, 1992, 32)

By Wilkes's lights, the task of SP is systematic explanation and description, and only certain kinds of phenomena are amenable to this project: natural kinds. Natural kinds, further, are things -- it is not clear that she is talking about properties, substances or other candidates -- that can figure into laws.

Certainly Wilkes owes us an account of the conception of law at work here. It is a part of many philosophical traditions that all things are held together by laws. Perhaps anticipating this, she claims that what makes a set a natural kind is that we can get *interesting* laws about its members. An account of an interesting law is therefore in order.

Such an account is certainly required if, as seems likely, a large part of her argument consists in the claim that conscious phenomena, as the vernacular has it, are not themselves a natural kind. She admits that '[m]uch evidently revolves around the "usefulness" and the "interest" of laws...'(Wilkes, 1988, 32) and deciding what counts as a natural kind 'require[s] us to have some grip on what counts as *interesting* or informative laws.' (Wilkes, 1988, 32)

Despite her acknowledgment of the importance of giving an account of what makes a law interesting, she says,

> But here I admit that I am issuing promissory notes, because of constraints of space; in general, to decide on what counts as an 'interesting' law is a fascinating and important problem I cannot discuss now. Full treatment would require a look at all the 'values of scientific theories', such as their scope, range, accuracy, simplicity or economy, predictive power, capacity for extension to new domains, internal logical consistency, consistency with theories in related areas, and so on and so forth....Here I must reluctantly leave the idea of 'an interesting law' to intuition. (Wilkes, 1988, 32)

Obviously, more than this is required, but it is a start, and no doubt our intuitions about laws will come into play in the rest of this chapter. For now, at minimum we can say that an interesting law is non-trivial and probably not *ad hoc*, but that is not much help. What matters most seems to be, for Wilkes and others, our scientific values: simplicity, scope, fruitfulness, consistency and the like. To be a kind, on Wilkes's view, a set has to be subject to laws that conform to intellectual virtues such as these.

What more does Wilkes claim about kinds? She distinguishes between more and less unitary kinds. Highly unitary kinds -- proper natural kinds, on her view -- are such

that all or virtually all the members of the kind are subject to the same laws. Less unitary kinds, or cluster kinds, fall into sub-classes, where a few laws apply to all or virtually all members of the kind and other laws apply only to constituent sub-classes or groupings of sub-classes. Cluster kinds hang together in virtue of a sort of structural isomorphism between the laws that govern its members, she claims. 'Tiger', 'gold' and water' are examples of natural kinds; and 'metal', 'acid', and 'fish' are examples of cluster kinds.

She contrasts such kinds with mere arbitrary sets, like words that begin with the letter 'g'. It is implausible, according to Wilkes, to suppose that we are likely to arrive at interesting laws covering mere arbitrary sets. 'Consciousness', on her view, is closer to the arbitrary set end of the spectrum, lumping together what she takes to be a hodge-podge.

There is little more in the way of clarification of natural kinds, cluster kinds, and arbitrary sets here. Clearly, very much depends on the notion of interesting law, which Wilkes leaves to our intuitions. Her intuitions, at least, are such that the things picked out by the vernacular conception of consciousness are not the sort of things one can hope to have interesting laws about.

An argument from the nature of CSP terms

Now we are in a position to clarify Wilkes's move from the many uses of 'conscious' to the conclusion that the sciences need not bother with consciousness. She writes,

> ...'consciousness' and 'conscious' are terms of the vernacular which not only need not but should not figure in the conceptual apparatus of psychology or the neurosciences, for the concepts required in those theories will want to group phenomena along different, more systematic principles. We can support this argument by looking briefly at the way we use the notion in our everyday language; for by so doing it will be seen that consciousness is not the sort of thing that has a 'nature' appropriate for scientific study. (Wilkes, 1984, 237)

The idea is this. 'Consciousness' and its cognates are terms of CSP. CSP has goals other than systematic description and explanation, and to pursue those goals, the language of CSP sometimes groups phenomena in an unsystematic fashion. It lumps things together for its own purposes, and the purposes of CSP do not often include the systematic description and explanation of kinds in terms of interesting laws.

A consideration of four everyday uses of 'conscious' issues in a number of conclusions we have already identified: 'consciousness' is a crude notion, ill-equipped to make

precise distinctions, picking out a hodge-podge, leaving us with only a few clear paradigm cases and nothing more, and so on. From this it follows, Wilkes claims, that 'consciousness' as the vernacular has it does not carve the world up such that it picks out a natural kind or even a cluster kind. Hence, it is not the sort of thing that has a nature amenable to systematic explanation and description in terms of laws, which is characteristic SP. If the importance of a term for SP consists in its potential suitability for systematic study, then 'consciousness' is not important for SP.

More formally, (i) 'consciousness' as used in CSP does not pick out a natural kind (given the goals of CSP, Wilkes's understanding of kinds, and the conclusions reached in her discussion of the uses of 'conscious'), (ii) only the referents of terms that pick out precisely natural or at least cluster kinds are appropriate for scientific study (given her account of SP and her understanding of kinds), (iii) therefore, 'consciousness' does not name appropriate subject matter for SP; and insofar as SP need deal only with appropriate subject matter, SP need not bother with 'consciousness'.

Now, the obvious trouble with this argument is that it seems to prove too much. The argument cannot just be that 'consciousness' is a CSP term, and the nature of CSP terms is such that SP need not bother with items picked out by them. This leaves us with the intolerable view that nothing we refer

to in our everyday talk is amenable to scientific study.

There seem to be scores of CSP terms that figure into work carried out by SP, despite the claim that CSP terms do not pick out bits of the world that are appropriate for systematic exploration. Think of 'memory', 'language', 'emotion', 'intelligence', 'seeing', 'hearing', 'smelling', 'tasting', 'touching', 'believing', 'imagining', and 'pain'. Worrying for Wilkes is the apparent fact that SP also makes use of 'attention', 'awareness', and 'consciousness'. A great deal of empirical work seems prefaced on the notion that subjects are conscious of some things and not others, can attend to this while being unaware of that, and so on. So as not to beg any questions, let us bracket the apparently successful use of 'conscious' in SP for now.

If the line Wilkes is taking is just that CSP terms do not pick out appropriate subjects for scientific study and 'consciousness' is a CSP term, then the argument seems to prove that SP need not and should not bother with any CSP term. That proves too much, even for Wilkes. After all, she is claiming only that 'consciousness' is unimportant, not that all CSP terms are unimportant. Yet the argument, as it stands right now, seems geared to prove this larger, exceptionally dubious claim. Further, the argument as it stands faces a number of counter-examples -- any of the mentioned recognizable CSP terms employed in SP.

Heterogeneity, imprecision, and second-order ascription

So what is needed is an amendment to the argument that makes 'consciousness' the focus, not the whole of CSP. There has to be something about this term in particular that makes it unimportant from the point of view of SP. Wilkes admits that some CSP terms, 'intelligence' and 'pain' for example, have been incorporated into SP. However, she says, such terms are different from 'conscious' in certain important respects. She argues as follows.

Science clearly adopts certain terms from CSP. Sometimes science simply borrows a word without regard for its everyday meaning -- 'charm', for example. On other occasions, though, an adopted term is only partially modified to meet the requirement of systematicity characteristic of SP -- fuzzy conceptual edges are merely smoothed over by the needs of rigour and exactitude, leaving a concept recognizable from the standpoint of common sense.

In the latter sorts of cases, Wilkes argues, a term is modified only a little, such that it makes sense to say that what science is talking about is roughly the same as what the average person is talking about. Wilkes maintains, for example, that 'intelligence' is operationally defined in SP in terms of success on IQ tests, and this seems to mesh more or less with the everyday use of the term. Other borrowed terms are modified more and more until we reach words like 'charm' which seem to have nothing in common with their common sense

origins.

Why does Wilkes claim that intelligence but not consciousness is amenable to scientific study? She takes up the question herself, arguing that the difference consists in two facts: 'conscious' is imprecise and names a heterogeneous set of things, and 'conscious' is a second-order term.

Consider heterogeneity and imprecision first. The familiar conclusions she reaches with respect to the four uses of 'conscious', she maintains, suggest that adoption by SP of the CSP term 'conscious' will require such revision that we will no longer be talking about anything like the common sense understanding of the thing. She writes,

> ...an immediate implication of [her consideration of the uses of 'conscious'] is that no precise characterization could come close to capturing the thoroughly *im*precise and heterogeneous everyday meaning that the term has in the vernacular. The adaptation that the term 'conscious' would need to undergo before it could be made to cover tidily a systematically related bunch of behaviours would be so great that a study of *this* 'consciousness' would no more be a study of consciousness as we think of it than the study of the spin of an electron can inform us about the behaviour of spinning-tops.

(Wilkes, 1984), 239)

She admits that some heterogeneous CSP terms can be modified such that SP can employ them, without too much divergence from the common sense use. 'Pain' is her example. She claims it is heterogeneous -- 'few sensations are less similar than a dull stomach ache and the prick of a pin' (Wilkes, 1984, 234) -- but a suitable modification of the CSP concept in terms of functional role might serve to unify the notion for scientific purposes.

However, 'consciousness' is different. She seems to argue that 'consciousness' (but, presumably not 'pain') has a meaning so imprecise and heterogeneous that massive adaptation would be required before it could serve as an SP term. The adapted term would be as changed as 'charm', 'spin' and other such terms at the far end of the modification spectrum.

Second, Wilkes claims that 'consciousness' is a second-order term, picking out a second-order property. A second-order property is ascribed only if certain first-order ascriptions are appropriate. To use her example, a person is called 'intelligent' if she performs a number of tasks well, with a certain degree of plasticity and persistence, skill and sophistication. IQ tests, she says, 'do not test intelligence, but rather test performance at a set of first-order tasks, and intelligence is ascribed to the extent that these are done well'. (Wilkes, 1984, 237) Since

'intelligence' is her example of a CSP term that has undergone smooth assimilation into SP, we can conclude at least that being a second-order term need not disbar a CSP word from guiding scientific study.

'Consciousness' too, on her view, is ascribed only if a number of other psychological predicates are appropriate. She writes,

> ...we presuppose a whole slew of psychological ascriptions -- to do with perception, motivation, belief and desire, misperception, illusion, recognition, *etc.* -- when an ascription of consciousness makes sense. (Wilkes, 1984, 238)

Presumably, she is talking about the first use of 'conscious' here, the use applied to whole conscious systems. If she is right, then the first use is a second-order predicate. It is not clear what she thinks about the other three uses.

Her discussion of why some CSP terms, but not 'conscious' is amenable to scientific study abruptly ends with the following dense conclusion:

> In sum, then, the fact that an ordinary-language term covers a heterogeneous range of phenomena need not bar it from study (in tidied-up form) by science -- 'pain' can be and is studied; the fact that a

> term is a second-order term, like 'intelligence', does not prevent the same treatment; but 'consciousness' is both at once, while compounding its intractability, unlike 'pain' and 'intelligence', by the further fact that its meaning is or seems to be *completely* exhausted by its use: no adequate truth-conditions can be imposed upon ascriptions of consciousness. (Wilkes, 1984, 239)

The claims made here are as follows. 'Conscious' is both heterogeneous and imprecise, and it is a second-order term. Being heterogeneous and imprecise dims the prospects for smooth assimilation into SP, because it suggests that 'conscious' does not pick out a natural kind.

Being a second-order term is bad for 'conscious' too, but the reason is never made explicit in text. The best answer seems to be that second-order ascriptions presuppose a number of first-order ascriptions, and identifying the ascription conditions of a number of first-order predicates can quickly become 'a highly complex business....'(Wilkes, 1984, 238) Worse for 'conscious', Wilkes claims the that first-order mental predicates which constitute the ascription conditions for 'conscious' do not themselves have ascription conditions. She writes,

> I cannot specify necessary and sufficient conditions

> for the applicability of 'see', 'hear', 'want', 'afraid' and so forth any more than I can for 'conscious', because I do not believe that there are any such. (Wilkes, 1984, 238)

This must be the reason for the final claim in the passage, that no adequate truth-conditions can be imposed upon ascriptions of 'conscious'. If 'conscious' is a second-order term, and the first order-terms required for its ascription do not have necessary and sufficient conditions for application, then, Wilkes concludes, there just are no truth-conditions for ascriptions of 'conscious'.

Final version of the argument

Now we are in a position to draw these various strands together and formulate a final version of Wilkes's argument, one focusing on consciousness, not the whole of CSP.

First of all, 'conscious' as used in CSP is not just heterogeneous and imprecise -- being so need not disbar a term from adoption by SP. To distinguish 'conscious' from other terms in CSP, it has to be *exceptionally* heterogeneous and imprecise, such that adoption would require a transformation that would leave the new term as different from its common sense incarnation as 'charm' in science is different from 'charm' in the vernacular.

That is an extremely ambitious claim that is worth

considering. The meaning of 'charm', used in science shares nothing with the meaning of the common sense term. Keeping the word 'conscious' in SP, then, would be merely cosmetic -- the new concept or concepts used would share as much as the scientific conception of 'charm' does with the common sense conception, which is to say nothing at all.

Second, 'conscious' is a second-order term. Again, this alleged fact alone is not enough to distinguish 'conscious' from other second-order CSP terms that have been adopted by SP. The difference for 'conscious', Wilkes needs to say, is that, unlike CSP terms such as 'intelligence', there are no necessary and sufficient conditions for the applicability of the first-order predicates required for the ascription of 'conscious'. If we cannot say with precision when someone 'sees', 'believes', 'hurts', and so on -- and we cannot, if there just are no truth-conditions for the ascription of such terms -- then we cannot ascribe 'conscious' with precision. If SP needs terms conforming to this conception of precise reference, then 'conscious' will not do for SP.

Thus, we have a reply to the difficulty with the original version of the argument. What makes 'conscious' different from other CSP terms such that it (but not all CSP terms) is unimportant from the standpoint of science? By Wilkes's lights, 'conscious' has more strikes against it than the average CSP term. It is excessively imprecise, picking out an extremely heterogeneous set of things. It is also a second-

order term with a particular defect: the ascription conditions for the first-order terms underwriting its ascription have no truth-conditions. 'Pain' and 'intelligence' are heterogeneous and second-order respectively, but they are not both at once. 'Conscious', according to Wilkes, is, and this renders it doubly intractable.

If all of this is so, then it is very unlikely that 'consciousness' and its cognates can be smoothly incorporated into SP. Thus, SP can ignore it. As Wilkes says, it does not pick out the right sort of phenomena for systematic study.

How does Wilkes move from this conclusion with respect to the utility of 'consciousness' for SP to the (admittedly speculative) claim that ordinary language and philosophy need not bother with the notion either? She maintains that 'the arguments sketched above, although designed primarily for the sciences, can be readily adapted to show that even in everyday discourse ["consciousness"] has little explanatory or descriptive value'. (Wilkes, 1984, 241)

Her thinking seems to be just that, insofar as CSP and philosophy of mind are interested in description and explanation (and CSP's interests here are limited, on Wilkes's view), both can dispense with the notion as well. The problems associated with heterogeneity and imprecision and the worries associated with the ascription conditions of the first-order predicates that underwrite ascriptions of 'conscious' are

problems for the utility of 'conscious' in philosophy and CSP only insofar as both engage in description and explanation. To the extent that philosophy and CSP engage in activity like SP, neither need bother with consciousness.

3.3. The Alleged Heterogeneous, Imprecise, and Second-Order Nature of 'Conscious'

Let us take up the three particular claims made against the utility of 'conscious' -- it is heterogeneous, imprecise, and second-order -- with a view towards determining whether they are true.

Heterogeneity

What is it to say that 'consciousness' names a heterogeneous set of things? A distinction between two senses of heterogeneity might help clarify Wilkes's position. There is a trivial sense in which all sets with more than one member are heterogeneous -- there is at least one property that differs among them. For any assertion of heterogeneity to be interesting it has to be made against the background of some relevant standards or criteria. Heterogeneity, in other words, is where you find it, and you can find it almost anywhere. However, interesting claims to heterogeneity are made with respect to some criteria relevant to the discussion.

To use Wilkes's example, 'tiger' names a natural kind. However, there is enough heterogeneity amongst tigers and the

interests of biologists are such that eight different sorts of tiger have been identified (South Chinese, Siberian, Indian, Sumatran, Indochinese, Javan, Caspian, and Bali tigers). No doubt there is some standard and some shared properties in virtue of which it makes sense to call them all 'tigers', and once all of this is known whatever heterogeneity remains is ignored for purposes of 'tiger' ascription. Pointing out that some tigers are smaller than others, have different dietary habits, social structures, fur thickness and colour, and so on is irrelevant to the general claim that all such creatures are tigers, given our interests and their properties. The relevance of claims about the heterogeneity of a set depends on at least two things: some standard and the way the world is.

It is no good saying that a class is heterogeneous if it is not clear what criteria are used to make the claim. It is also no good if the criteria are irrelevant to the discussion at hand. So what criteria are used in Wilkes's claim that consciousness is heterogeneous? Wilkes's efforts seem geared to show just that the four things called 'conscious' are distinguishable from one another -- she seems to aim for demonstrating heterogeneity in the trivial sense. She spends considerable time distinguishing between sensations and perceptions, and also maintains that both the attitudes and sensations are themselves heterogeneous -- in short, simply showing that there are differences between things that are

called 'conscious'. As we have seen with the set of tigers, identifiable differences alone do not establish that the set itself is not homogeneous in an interesting sense.

Perhaps she is operating on the assumption that the more differences identified, the less likely it will be that interesting similarities exist between the members of the set, similarities that might lead us to hope that interesting laws range over the members of the set. Perhaps this is behind her claim that the set of things called 'conscious' is grossly heterogeneous. This strategy does not prove very much. Unless we know what the criteria are, a list of differences -- even a very long list -- is no help in determining whether or not a class is interestingly heterogeneous.

It might be said that Wilkes's criterion is 'being fruitful *explananda*' or, what amounts to the same thing for her, 'being held together by interesting laws'. This line does not seem to be open to her. As we have seen, she argues from a discussion of general heterogeneity, a list of differences between things called 'conscious', to the claim that the members of the set are not fruitful *explananda*, not likely to be held together by interesting laws, and on to the claim that SP need not both with consciousness. Yet, however many differences one might identify amongst the members of a set, it is a *non sequitur* to conclude from this that no interesting laws bind its members together. There might still be one or more properties, as yet unidentified, that renders

the set interestingly homogeneous. In fact, it seems likely that all conscious states are interestingly homogeneous with respect to at least one property. An example will help make the point.

Consider the adjective 'triangular'. Following Wilkes's thinking, one could formulate an argument for the uselessness of 'triangularity' for science by arguing that 'triangular' is applied to a heterogeneous set of things. It is clear that there are very many differences between things like wedges of cheese, aeroplane wings, coat hangers and door stops, yet they are all triangular things. Does this show that 'triangular' names a heterogeneous set of things? In the trivial sense, it certainly does, but triangular things still share certain properties in virtue of which it makes sense to group them together for certain purposes. Despite trivial heterogeneity, 'triangularity' is still a useful concept for many scientific purposes, from pure maths to the application of certain principles of aviation.

'Conscious' might be applied to a heterogeneous set of things in the same way that 'triangular' is. Nevertheless, there might be certain properties shared by all conscious perceptions, sensations, and propositional attitudes in virtue of which the concept might be useful for scientific purposes. According to an everyday conception of consciousness, it is clear that such states sometimes share one very interesting property: we are conscious of them. Just as triangular

things are homogeneous with respect to the standard of triangularity, states such as these are homogeneous with respect to the property of being states of consciousness.

It is true that considerable clarification is needed if we are to come to a real understanding of this property; however, in a simple-minded way it is clear enough that, to use a now well-worn phrase, there is something it is like to have perceptions, sensations, and attitudes. Whatever heterogeneity attends such states seems to be swamped by this simple fact, just as the differences amongst triangular things are put aside when we reflect on their triangularity.

When Wilkes asks, 'what could be less similar than a dull stomach ache and the prick of a pin', it seems reasonable to reply that having a dull stomach ache and being a rock are considerably less similar, if what interests us is the way things feel. There is something it is like to have a dull stomach ache or feel a pin prick, but no such property attends rocks, pricked by a pin or otherwise. This property, being conscious of something, binds the class of conscious things together in what seems like a robust and interesting way, distinguishing conscious states from other states, like being round. My intuitions, at least, are such that we have here the kind of phenomenon about which it makes sense to think we can get interesting laws. For these reasons, Wilkes's claim that 'consciousness' names a heterogeneous set of things seems either trivial or mistaken.

Imprecision

There are several difficulties associated with Wilkes's claims about the imprecision of 'conscious' as well. We will consider three of them.

First, there might be an ongoing ambiguity in at least a part of Wilkes's discussion of imprecision that is best expressed by the following two claims: (1) 'consciousness' cannot be used to discriminate between different conscious states, and (2) 'consciousness' cannot be used to discriminate between conscious states and un- or non-conscious states. Claim (1) seems to be a trivial truth which attends all general concepts. Wilkes needs to establish (2) for her argument to go through, but she spends considerable time arguing for (1).

Consider this distinction with respect to her claims about the first use of 'conscious'. For the claim that 'conscious' is imprecise to go through, Wilkes needs to show that 'conscious' cannot be used to distinguish between people who are conscious and other things. However, on one reading, she actually argues for the claim that the general concept of the unconscious cannot be used to distinguish between the various stages of REM and NREM sleep -- in short, things that are in or a part of the unconscious. The pattern is regularly repeated in her discussion of the other uses of 'conscious' as well.

Of course a general concept is no use when trying to set

off distinctions within it. What general concept is useful in such a context? Consider 'tiger' again. It is no argument for the imprecision of 'tiger' to point out that the general term cannot be used to distinguish between South Chinese and Siberian tigers. Research into the differences between different sorts of tiger will require something other than the general concept. So too with consciousness: research into the differences between different conscious states and different un- or non-conscious states will require something other than general concepts. This does not point to the imprecision of 'consciousness', but to a simple fact about the nature of general terms.

If one wants to show that 'tiger' is imprecise, one would need to show that the concept cannot be used to distinguish between tigers and other things. What Wilkes needs to do is show that 'conscious' in the first sense cannot be used to distinguish between conscious people and other things. Instead she argues that 'conscious' cannot be used to distinguish between different states that are in consciousness, and 'unconscious' cannot be used to distinguish between the different states that are part of the unconscious. The concepts cannot cope with all this diversity; something more precise is needed, she argues. This is plainly arguing for the wrong thing.

However, at least some of her claims do aim towards establishing that 'conscious' cannot be used to distinguish

between states which are conscious and states which are un- or non-conscious. Unnoticed pains are a good example. Should they count as conscious or unconscious states? Wilkes believes that such pains are neither clearly conscious nor clearly unconscious; the imprecise concepts at our disposal leave unnoticed pains uncharacterized. Scientific psychology, Wilkes concludes, cannot proceed on the basis of such imprecision and would do well not to bother with consciousness.

One might respond to this view by claiming that, though our current use of 'conscious' leaves us with problem cases -- perhaps not as many cases as Wilkes maintains, but problem cases nonetheless -- this fact alone does not compel us to accept the conclusion that the word cannot be further clarified. Wilkes's response is that imprecision is essential to 'consciousness' and its cognates -- rooting out the imprecision would leave us with a radically different concept, one not recognizable from the common-sense point of view, much like 'charm' or 'spin'. It is worth wondering, if only for a moment, if there are any other terms in the history of philosophy with the kind of essential imprecision Wilkes is crediting to 'consciousness'. It is not clear that there are.

Wilkes's stance here suggests a second difficulty. The vernacular conception of consciousness seems considerably more plastic than Wilkes's line allows. The common-sense conception, in other words, seems amenable to clarification in

virtue of input from science, and this suggests that the imprecision Wilkes posits is not essential to the concept. Wilkes acknowledges this, if only tacitly, in her consideration of sensory perception.

In an effort to build a case for the heterogeneity of consciousness, she argues that sensations are different from perceptions because, for sensations but not perceptions, to be is to be perceived. Given the results of recent studies of phenomena like blindsight and subliminal perception, we now admit that some instances of sensory experience are conscious and others non- or un-conscious.

This seems like a very clear case of common-sense talk being rendered more precise in virtue of scientific discovery. If this sort of thing can happen, and Wilkes herself admits that it can, then what stands in the way of further clarifications, eventually issuing in a conception of consciousness precise enough for serious empirical and philosophical purposes, yet recognizable from the common-sense point of view? The modification resulting from the study of blindsight and subliminal perception is more precise than its conceptual forebear, and it is still a recognizable conception of consciousness. Why not more progress in this direction?

Most of what Wilkes says about imprecision seems to support the conclusion that we do not have a precise conception of consciousness -- this is a conclusion that few would deny. The claim she wants to establish, though, is not

that we do not have an adequate conception, but that we cannot get one. In order to move from the inadequacy of the current conception of consciousness to the claim that no future modification of this conception will have the precision necessary for serious study, Wilkes builds essential imprecision into the current concept of consciousness. This presupposes just what Wilkes needs to prove. It also seems to conflict with what Wilkes admits elsewhere, that the everyday conception can be rendered more precise without ending up like, 'charm', 'spin', and the like.

There is a third and final worry worth considering. Wilkes operates with the assumption (more or less followed in this chapter) that facts about the adjective 'conscious' have certain implications for the nature of the noun 'consciousness'. In particular, the difficulties associated with the ascription of 'conscious' suggest to Wilkes that we cannot ever arrive at a clear conception of 'consciousness'. This move from difficulties with the adjective to difficulties with the noun might not be legitimate, as the following example makes plain.

We can say with great clarity what a muscle is -- a collection of cells that can contract -- even though 'muscular' is an imprecise adjective that admits of many problem cases. Some individuals suffering from lateral paralysis have well developed muscles on only half their bodies, since those muscles are often doubly exercised.

Perhaps it is difficult to say whether or not such individuals are muscular. It does not follow from our difficulties with the adjective that an adequate conception of 'muscle' cannot be had or that 'muscle' has no place in our theorizing about body mechanics. Analogously, it does not follow from problems associated with the ascription of 'conscious' that an adequate conception of 'consciousness' cannot be had or that 'consciousness' cannot inform theorizing about the mind.

For these three reasons, it seems prudent to reject Wilkes's claims about the imprecision of 'conscious'.

'Consciousness' as a second-order term

The bulk of Wilkes's arguments attempt to establish the claim that 'conscious' is heterogeneous and imprecise. The main argument, as we have seen, takes this claim and couples it with the further claim that 'conscious' is a second-order term. Being heterogeneous and imprecise or second-order does not prevent adoption by SP, as in the cases of 'pain' and 'intelligence'. The problem for 'consciousness', Wilkes argues, is that '"consciousness" is both [heterogeneous and imprecise *and* second-order] at once' (Wilkes, 1984, 239).

On one reading, this looks like equivocation. The arguments for heterogeneity and imprecision are aimed at 'conscious' in the sense of states of or in consciousness -- conscious bodily sensations, sensory experiences, and propositional attitudes. The claim that 'conscious' is

second-order is aimed at 'conscious' meaning being awake.

Her main line of thinking can be paraphrased as follows.

(i) 'Conscious' picks out a heterogeneous and imprecise set of things (bodily sensations, sensory experiences, and propositional attitudes).

(ii) 'Conscious' (being awake) is second-order.

(iii) Being heterogeneous and imprecise or second-order need not disbar a term from adoption by SP.

(iv) 'Conscious', however, is all three at once.

(v) Therefore, SP need not bother with it.

'Conscious' in (i) means states of or in consciousness, like visual experiences and beliefs. 'Conscious' in (ii) means being awake, having the sorts of inner states just mentioned. If we accept (i) and (ii) in the sense Wilkes intends, then it is difficult to see how (iv) could be true.

'Conscious' is being used in two very different ways in the argument. It might be thought that Wilkes's point is that the difficulty is not hers, but a problem with the concept of consciousness itself. We are so confused in our thinking about consciousness that we regularly equivocate, and her argument just draws attention to this. Despite the fact that there is no textual support for this reading, so far as I can tell, it seems a little outrageous to maintain that our thinking of consciousness is so confused that we cannot

discriminate between people who are conscious and the conscious states that they have. How likely is it that we regularly mistake beliefs or bodily sensations for conscious people?

There is another way to understand Wilkes here that saves her from the charge of equivocation. The claim that 'conscious' is heterogeneous and imprecise is to be understood as directed towards states in consciousness like the ones Wilkes considers. The claim that 'conscious' is second-order is directed towards 'conscious' meaning people who are awake. On this reading, the second claim is dependent on the first one.

In other words, Wilkes argues as follows. 'Conscious' as applied to sensations, experiences and attitudes is heterogeneous and imprecise. 'Conscious' as applied to people who are awake depends for its ascription on 'conscious' as applied to sensations, experiences and attitudes -- we do not call someone 'conscious' (awake) unless he can enter into these sorts of first-order states. Because the first-order use of 'conscious' is imprecise and names a heterogeneous set, no truth-conditions can be found for the second-order use. 'Conscious', on either use, cannot be much help to SP.

Several problems present themselves. First, as we have seen, there are good reasons for denying Wilkes's claims about heterogeneity and imprecision with respect to first-order ascriptions of 'conscious'. If her arguments concerning

'conscious' as a second-order term depend on these earlier claims, then the suspicions voiced earlier are applicable here as well.

Second, by escaping the problem of equivocation, Wilkes is no longer entitled to the claim that 'conscious' is heterogeneous and imprecise and second-order at the same time. Given that she admits that being either heterogeneous and imprecise or second-order need not disbar a term from adoption by SP, this move leaves her with no argument for her conclusion, that SP need not bother with consciousness.

There is a final worry for Wilkes here. No one is called 'intelligent', Wilkes argues, unless he or she can execute a range of different tasks with a certain degree of skill, flexibility and originality. Clearly, if a person can do only a few of the relevant things, then, by Wilkes's lights, the person is not intelligent. If 'consciousness' really is a second-order term like 'intelligence', then a similar relationship ought to hold between the ascription of certain first-order predicates and 'consciousness'.

The point is speculative, but it might be argued that ascriptions of 'conscious' do not depend on the ascription of a number of first-order predicates in a manner analogous to 'intelligence'. One probably harbors doubts about crediting intelligence to an autistic savant capable of rapid calculations but not the other skills needed for the ascription of the first-order predicates that underpin

intelligence. However, the same sort of doubts do not immediately arise if we imagine a person sitting alone in a darkened room, meditating on the number four, or a locked-in patient, of whom one can fairly ascribe only one or two of the first-order predicates under consideration. Though very many of the first-order terms needed for the ascription of 'conscious' cannot be predicated of either person, both seem no less conscious. The intuitions we have about such cases suggest that 'conscious' might not be a second-order predicate at all.

Conclusions

Wilkes's case against 'consciousness' consists in the claim that the vernacular use of the word suffers from certain defects that make it an unlikely tool for systematic study. Our consideration of this view results in three principle conclusions.

First, claims that 'conscious' names a heterogeneous set of things requires some standard or criterion with which one might judge the truth of the claim. Wilkes provides none. This alone is enough to make it possible that the set in question admits of interesting homogeneity with respect to some as yet undiscovered property. Reflection on the intuitive difference between states called 'conscious' and other states, like being round, suggests that some property really does serve to unify conscious states, though we have an

imperfect grasp of its nature. Like so many questions in this neighborhood, it is likely that the matter will be decided empirically, by continuing investigation into whatever we are getting at with locutions like 'phenomenal feel', 'what it's like', 'subjectivity', and so on.

Second, 'conscious' as used in the vernacular certainly leaves us with problem cases, but as we have seen not much about the possibility of coming to a better conception of consciousness follows from this -- unless we are persuaded by the possibility of essential imprecision built in to any concept of consciousness recognizable as such. As Wilkes appears to admit, the everyday concept of consciousness has already undergone modification as a result of empirical and conceptual work. There is no clear reason to believe that this process cannot continue, resulting in a conception of consciousness precise enough for serious scientific and philosophical purposes.

Third, the view that 'conscious' is second-order seems applicable, if at all, to the word in so far as it means 'being awake'. This fact undermines Wilkes's claim that the intractability of 'conscious' consists in its being heterogeneous, imprecise and second-order at once. As we have seen, it is arguable that 'conscious' is not second-order at all.

In the end, all of this points to something that we probably knew already: the everyday conception of 'conscious'

is, for whatever reason, not adequate for a certain kind of philosophical and scientific work. There are some questions that we cannot answer, perhaps even formulate, because of our imperfect grasp of consciousness. We are in a position, though, to ask more basic questions, and it is a mistake to believe that progress here requires the kind of conceptual clarity needed further down the road. It seems likely that much science, and no little amount of philosophy, just is tinkering around with things imperfectly understood. Perhaps this is how things stand with consciousness. Nevertheless, we do not yet have good reason for thinking that consciousness is not the sort of thing amenable to scientific or philosophical study.

4. Evolution and Understanding Consciousness

We have considered arguments representative of the eliminativist camp, and one argument for the claim that consciousness is not important, and so far we have discovered no good reason for thinking that consciousness cannot or need not be understood. We turn now to reasons that might be given for the claim that an understanding of consciousness is somehow beyond us. According to the view considered in this chapter, consciousness, real enough and important to our thinking about the mind, cannot be understood because the problem solving capacities at our disposal are circumscribed by certain evolutionary needs, needs that have nothing to do with understanding consciousness.

The view begins in reflection on the nature of human understanding in so far as it is a product of evolutionary forces. The faculties we bring to bear on the questions that concern us -- what is consciousness and how is it related to the rest of the world -- are designed features of human beings. All such features are what they are in virtue of selection pressures, and, therefore, our faculties are constrained by the need to understand what is necessary to ensure reproduction and survival. Answering the questions of interest to us, it might be argued, is of little consequence to our evolutionary needs. It seems unlikely, then, that evolution has equipped us to understand consciousness.

The claim that the qualitative aspect of conscious states

cannot be understood, given the effects of evolutionary forces on our capacity to understand, appears in the work of Jackson. (Jackson, 1992, 475) The claim is embedded in a larger project, an attack on physicalism, and it is part of an anti-physicalist view that Jackson no longer holds.(Jackson, 1998) Because our interest is whether or not there are good reasons for thinking that consciousness cannot be understood, we will consider the argument anyway, even though Jackson's current views might be some distance from his earlier thoughts. The conclusion is of interest, whether or not Jackson now defends it.

The argument from evolutionary considerations is a response to a general objection to Jackson's claim that the qualitative aspect of conscious states is epiphenomenal, an objection that surfaces only after an extended consideration of three very specific objections to epiphenomenalism. It might be unwise to ignore this background, for three reasons.

First, a number of remarks about evolution are made prior to the general argument from evolutionary considerations, remarks that will become very relevant in our appraisal of the argument. Second, the setting of the argument gives us a number of clues concerning the scope of Jackson's skepticism -- the sense in which we might not be able to understand consciousness, in other words -- and this is worth considering if we are to have a grasp of his version of the claim that consciousness cannot be understood. Third, in the light of

the setting of the argument from evolution, there might be grounds for concluding that Jackson is, in fact, not arguing for the view that consciousness cannot be understood. If this is the case, then we have one less argument to worry about. The background, then, warrants consideration.

The dialectic is a complex one, so a little compartmentalisation is necessary. We begin with a consideration of his knowledge argument against physicalism. The epiphenomenalist position he takes up as a result of it raises three objections, and we will consider them, along with his replies, followed by the main argument from evolution. We conclude with a discussion of a certain kind of fallacy, which might be at the heart of Jackson's argument and similar arguments beginning with evolutionary claims and issuing in the conclusion that consciousness cannot be understood.

4.1 Epiphenomenal Qualia

Jackson's discussion begins with a general attack on physicalism. His main argument makes use of two notions: physical information and qualia. Both notions are extremely difficult to pin down, and Jackson's account of each leaves a number of interpretive questions open. We shall consider each in turn, for the notions provide a backdrop to both his general argument against physicalism and, more importantly for us, his claims about understanding consciousness.

Physical information and qualia

Jackson does not claim to provide a definition of physical information, but he does try to characterize it with the following, admittedly sketchy remarks:

> It is undeniable that the physical, chemical and biological sciences have provided a great deal of information about the world we live in and about ourselves. I will use the label 'physical information' for this kind of information, and also for the information that automatically comes along with it. For example, if a medical scientist tells me enough about the processes that go on in my nervous system, and about how they relate to happenings in the world around me, to what has happened in the past and is likely to happen in the future, to what happens to other similar and dissimilar organisms, and the like, he or she tells me -- if I am clever enough to fit it together appropriately -- about what is often called the functional role of those states in me (and in organisms in general in similar cases). This information and its kin, I also label 'physical'. (Jackson, 1992, 469)

The first part of the definition ties the nature of the

physical to current theories at work in physics, chemistry and biology. This might be unwise, because it makes the nature of the physical dependent upon the truth of current theories and only on the truth of current theories. If our current theories turn out to be false, then we end up with no conception of the physical.

It might be thought reasonable to reply that current theories are the best we have at the moment, and using them as a basis for a conception of the physical is therefore the best we can do. This is certainly true. However, this way of conceiving of the physical does not really tell us what it is about the things that figure into our theories that makes them physical things. We have only examples of the physical, no real conception of the physical, on this view.

The second part of the definition extends the physical to what can be inferred if we are 'clever enough' in putting the pieces together -- information about, among other things, the functional roles of states of the nervous system. This too might be unwise, because Jackson seems to be claiming that physical information includes not just what the sciences tell us about mental states, but also what might be inferred from all of that.

As we shall see, one of Jackson's conclusions is that qualia are not captured by physical information. If 'physical information' includes not just what the sciences tell us about the brain, but also what we can figure out on the basis of

this, applying whatever new conceptions that might arise in our theorizing about the mind, Jackson's conclusion is very strong indeed.

The second notion that figures into his main argument is qualia. Here is his characterization of qualia:

> I am what is sometimes know as a 'qualia freak'. I think that there are certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes...[for example] the hurtfulness of pains, the itchiness of itches, pangs of jealousy...the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise [and] seeing the sky. (Jackson, 1992, 127)

So, on his view, qualia are features of bodily sensations, perceptual experiences, and emotions. We cannot gather from this whether or not qualia are features of just the sorts of mental states listed, whether or not such states have features in addition to qualia, or even whether or not some of the states listed, like some bodily sensations, might lack qualitative properties. We also do not know whether Jackson understands consciousness as consisting just in qualitative states or in something more.

Perhaps the reason for this thin characterization is that

all Jackson needs, for his argument to work, is just that there are at least some mental states with qualitative properties. Whether qualia attend all or some of the states listed is irrelevant, for his purposes. The thin characterization might also result from a certain implicit belief Jackson and others have about the obvious nature of qualia. It is more or less clear, though, that Jackson is talking about certain properties of some experiences when he talks about qualia.

The Knowledge Argument

Though mental states may have features in addition to the qualitative ones, on Jackson's view, qualitative properties are the elusive and interesting properties of mental states, he maintains, for this reason: no amount of physical information can capture them. It would be helpful to have a clear understanding of what he means by 'capture' here.

Assemble all the physical facts concerning the states of a human nervous system and the causal relations it bears to its environment and so on, and, Jackson argues, no matter how clever one is in fitting all of those facts together, the qualitative features of conscious experience will be left out. If physicalism entails the thesis that all facts are physical facts, then, he concludes, physicalism is false, because it leaves out qualitative facts, facts about how the states are for the person who has them.

Physical information, then, is not about or does not take account of, the qualitative properties of mental states, on Jackson's view. How much of consciousness is not captured by the physical facts, according to Jackson? It is difficult to glean a substantive answer to this question from what he says. Though he does not make the claim, it seems consistent with what he says that consciousness includes something more than qualitative properties. Consciousness, on an ordinary understanding of the thing anyway, includes more than the mental states Jackson lists (bodily sensations, perceptual experiences and emotions).

It might be claimed that propositional attitudes, mental imagery, and, perhaps, a sense of self are parts or aspects of consciousness that Jackson neglects here. It may be that one or another neglected aspect of consciousness has qualitative properties, but nothing Jackson says commits him to this. All that we can confidently conclude is that the qualitative properties of at least some aspects of consciousness -- some bodily sensations, perceptual experiences, and emotions -- are left out of the physicalist account.

The line of thinking Jackson is pursuing here, supplemented with thought experiments, is called the knowledge argument. Consider this familiar version of it. A scientist called Mary investigates, among other relevant things, neuroscience, from a completely black and white environment with the aid of a black and white television. Suppose that

she specializes in the neurophysiology of vision and acquires all the physical information there is about what goes on in the brain when a person views a ripe tomato. She learns everything there is to know about the wavelengths of light reflected by various objects, retinal responses, the effects that ensue in the nervous system, and the like -- in short, everything there is to know from the standpoint of physicalism about seeing red.

Now, Jackson asks:

> What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false. (Jackson, 1992, 471)

The problem for most physicalists is to explain the difference in Mary after she sees red in a way that does not make a commitment to nonphysical properties of consciousness.

Some physicalists, Dennett (1991, 398) for example, just deny that there is a before and after difference -- if she actually knew all the physical facts, and it is difficult to

trust our intuitions here, then seeing red for the first time would not constitute new knowledge for her. She would already know what it is like, or so Dennett argues.

Other physicalists (Lewis, 1991; Nemirow, 1980), however, admit that Mary learns something but deny that the knowledge acquired is propositional. This move might be understood in terms of a distinction between 'knowing that' something is the case and 'knowing how' to do something. On this view, Mary gains certain abilities or know-how: it might be that she can now recreate the experience of red in her memory and imagination (Nemirow, 1990), or perhaps what she can do is identify another red experience as the same experience via introspection, once a kind of physical template has been made in virtue of actually viewing a red object. (Lewis, 1991)

Others (Horgan, 1984) maintain that Mary does gain a kind knowledge that -- she knows now, for instance, that seeing red is exciting. However, the new knowledge is at the level of sense, not reference. The knowledge argument does not show that there is an aspect of the world overlooked by physicalism; it only shows that Mary can think about something she knew about before under a different concept.

I briefly mention these objections to cordon them off from another one that is of primary concern to us. It is not the knowledge argument itself that is of immediate interest (though we will have something to say about it in the final section of this chapter), but we need to consider it in order

to understand the moves that follow. Once the objections and replies that Jackson considers in response to the knowledge argument are before us, we arrive at a general objection to the epiphenomenalist position that Jackson takes up. It is his response to it that concerns us, for it is here that Jackson makes a certain claim about the prospects for coming to an understanding of consciousness. So let us begin with a consideration of Jackson's version of epiphenomenalism, the objections he anticipates, and his replies to them.

Epiphenomenalism: three objections and replies

Jackson's epiphenomenalist position emerges as he attempts to address the following objection to the knowledge argument: if qualia exist but are not physical, they seem nevertheless to

> ...have to be given a causal role with respect to the physical world and especially the brain, and it is hard to do this without sounding like someone who believes in fairies. (Jackson, 1992, 470)

One version of the problem is this: if one accepts the existence of conscious properties that are left out of a physicalist account, one seems burdened with spooky properties, properties that seem to have a place in the causal scheme of things but nevertheless get left out of physical

accounts. Believing in the existence of conscious properties then is on a par with believing in fairies. His response is to argue that qualia are epiphenomenal: we need not worry about how qualitative properties affect the physical because qualia have no effects on the physical.

The doctrine of epiphenomenalism takes many forms, and Jackson does try to distinguish the version he is adopting from others. He writes:

> Is there any really *good* reason for refusing to countenance the idea that qualia are causally impotent with respect to the physical world? I will argue for the answer no, but in doing this I will say nothing about two views associated with the classical epiphenomenalist position. The first is that mental *states* are inefficacious with respect to the physical world. All I will be concerned to defend is that it is possible to hold that certain *properties* of certain mental states, namely those I've called qualia, are such that their possession or absence makes no difference to the physical world. The second is that the mental is *totally* causally inefficacious. For all I will say it may be that you have to hold that the instantiation of *qualia* makes a difference to *other mental states* though not to anything physical. Indeed general

considerations to do with how you could come to be aware of the instantiation of qualia suggest such a position. (Jackson, 1992, 474)

There is quite a lot going on in this passage. The two clear commitments made with respect to qualia are: (1) the qualitative aspect of certain mental states has no physical effects, and (2) the qualitative aspect of certain mental states (probably) has effects on other mental states. There are a number of other suggestions made here and in Jackson's general defence of epiphenomenalism, and it might be best to represent his claims with a table, in the interests of clarity.

	Have physical causes?	Have physical effects?	Have mental causes?	Have mental effects?
Qualitative aspect of mental states	yes	no	?	probably
Other aspects of mental states	yes	possibly	?	?

We know from the things Jackson says here and elsewhere that he believes that '[q]ualia cause nothing physical but are caused by something physical' (Jackson, 1992, 475). It also

seems clear that he believes that other, nonqualitative aspects of mental states can have physical causes and, possibly, physical effects. He also makes it clear, in the long passage above, that his understanding of epiphenomenalism commits him to the view that qualia have no physical effects. He does claim, though, that qualia probably have effects on other mental states -- otherwise it would be difficult to understand how we come to know of them.

This passage also enables us to say a little more about the nature of the mental on Jackson's view. It suggests that some mental states have both qualitative and nonqualitative aspects. Exactly what the nonqualitative aspect of mental states is is not made clear, but it is tempting to suppose, given what he says elsewhere, that the nonqualitative aspect of mental states is included in the physical information (for example, the functional role of a state) we might gain when we investigate something like pain. It seems likely, then, that the nonqualitative aspect of mental states has both physical causes and effects, on Jackson's view.

It is worth remembering that this position is partly motivated by a version of the interaction objection. If it is to work, then the effects that qualia have on other mental states must be effects that they have on other, nonphysical aspects of mental states, presumably other qualitative properties. Otherwise we have the same problem of interaction between qualia and the physical.

However, Jackson claims that we come to know about qualia -- that is form beliefs about qualia -- in virtue of some causal relation holding between qualia and mental states that are not obviously qualitative, like beliefs. My belief that my back is killing me is not, in itself, obviously qualitative by Jackson's understanding of qualia as features of bodily sensations, perceptions and emotions. No doubt the pain in my back has qualitative properties, and it seems reasonable to suppose that those properties stand in some causal relation to my belief that my back is killing me, but the belief itself is not obviously qualitative in the way that the pain is.

To be charitable, there might be something about my belief that could be characterized as qualitative, but it is not obvious that this is all there is to my belief, and it is not obvious that the qualitative aspect of my belief (whatever that might be) is what stands in a causal relation to pain. It might be argued, then, that it looks like qualitative properties (the way back pain feels) stand in causal relation to mental states that are not obviously qualitative (a belief about back pain). Either Jackson goes epiphenomenalist about all mental states that stand in causal connection to the qualitative ones, or he still might have a problem with interaction.

Jackson expands his claims when considering objections to epiphenomenalism, and perhaps a consideration of this treatment will shed more light on his view. It might be best,

therefore, to consider those objections and his replies. This will also get us closer to the reply that concerns the prospects for coming to an understanding of consciousness.

The first objection is that it seems more or less obvious that the hurtfulness of pain is what causes people to avoid pain, say that pain hurts, and possibly have the belief that pain hurts, and so on. However, Jackson says, 'to reverse Hume, anything can fail to cause anything.' (Jackson, 1992, 474) Regardless of the obviousness of the connection, it might be that the qualitative aspect of pain and things like pain avoidance behaviour are both the effects of some underlying brain event and not actually causally correlated with one another. Though it seems to us that pains cause certain behaviour, this seeming 'is simply a consequence of the fact that certain happenings in the brain cause both' (Jackson, 1992, 474) behaviour and the associated qualia. Jackson writes:

> No matter how often B follows A, and no matter how initially obvious the causality of the connection seems, the hypothesis that A causes B can be overturned by an over-arching theory which shows the two as distinct effects of a common underlying causal process. (Jackson, 1992, 474)

The second general objection he considers goes as

follows. According to evolutionary theory, the traits we have are those which evolved, and evolved traits are conducive to survival. Presumably, qualia have evolved too, and so they should confer some survival advantage. Therefore, qualia cannot be epiphenomenal -- they must have some effects in the physical world that confer some survival advantage, otherwise they would not have been selected.

Because of its importance later, Jackson's reply is worth quoting in full:

> Polar bears have particularly thick, warm coats, The Theory of Evolution explains this (we suppose) by pointing out that having a thick warm coat is conducive to survival in the Arctic. But having a thick warm coat goes along with having a heavy coat, and having a heavy coat is not conducive to survival. It slows the animal down.
>
> Does this mean that we have refuted Darwin because we have found an evolved trait -- having a heavy coat -- which is not conducive to survival? Clearly not. Having a heavy coat is an unavoidable concomitant of having a warm coat...and the advantages for survival of having a warm coat outweighed the disadvantages of having a heavy one. The point is that all we can extract from Darwin's theory is that we should expect any evolved

> characteristic to be either conducive to survival or a by-product of one that is so conducive. The epiphenomenalist holds that qualia fall into the latter category. They are a by-product of certain brain processes that are highly conducive to survival. (Jackson, 1992, 474)

The idea is this: if it is claimed that qualia must have physical effects because they are selected traits, and all selected traits have some effects -- namely effects that are conducive to survival -- it is open to the epiphenomenalist to point out that this objection rests on a misunderstanding of evolutionary theory, a kind of adaptationism. It is not true that all traits are conducive to survival. Some traits, like the heaviness of polar bear coats, are incidental but nevertheless necessary concomitants of traits that have survival value. Though qualia have no physical effects, they are a by-product of brain events that have effects which somehow confer survival advantage. It is worth postponing a consideration of Jackson's thinking here until part three of this chapter, when other aspects of his position are made clear.

The third objection to epiphenomenalism that Jackson considers is rooted in our knowledge of other minds. On one understanding of the problem, we know about the minds of others based on an inference from their behaviour. People act

such that positing inner states is unavoidable, and among the inner states posited are qualitative ones. How could that inference work, the objection goes, if it were true that qualitative states had no physical effects? If none of a person's behaviour is the outcome of qualia, then we can never know that other people experience qualitative states.

Jackson's reply consists in arguing from one event (behaviour) back to its cause (a brain event) and out again to a different effect of that same cause (qualia). Jackson uses the following example to explain this chain of reasoning. If one reads in *The Times* that Spurs won, one has good evidence for believing that *The Telegraph* also reports that Spurs won, even though *The Telegraph's* report was not a causal outcome of the report in *The Times*. One here reasons from the report in *The Times* back to its cause, Spurs winning, and from that cause out again to a different effect, the report in *The Telegraph*. Analogously, one might reason from behaviour back to a certain event in the brain that caused it and then back out again to a different effect of the brain event, qualitative states.

It is worth bearing in mind that these replies only work, if they do work, in the context of the knowledge argument. That is to say, given the conclusion that qualia are epiphenomenal: (1) strong intuitions concerning the apparent causal connection between qualia and certain physical events can be explained away; (2) the existence of qualia can be

brought into harmony with Darwin by thinking of them as by-products of other properties that are conducive to survival; and (3) an untenable view of other minds can be avoided with the common cause hypothesis.

The individual claims made in each reply are not, on their own, particularly convincing. For example, the mere logical possibility, if that is what it is, that an underlying brain event actually causes both qualia and associated behaviour is not enough, on its own, to convince us that the way pain feels is causally inefficacious. However, given the knowledge argument, the common cause hypothesis explains our intuitions. So much hangs on the knowledge argument itself, and we will return to it in a moment. Now, however, we are in a position to consider the argument of central concern for us.

4.2 The argument from Evolution

Jackson admits that one might be tempted to respond to all three of his replies in the following way. Perhaps it is true that there is no obvious, knock-down objection to epiphenomenalism that precludes response. It may be possible to believe that anything can cause anything, that qualia are an inexplicable by-product of selected traits, and even that one can follow an inference chain from behaviour to brain events to qualia and thus have some response to the problem of other minds. Nevertheless, the great difficulty associated with taking up such a position is that one is left with

considerable epistemic angst over the whys and wherefores of qualia. As Jackson puts this general objection:

> ...[qualia] do nothing, they explain nothing, they serve merely to soothe the intuitions of dualists....In short, we do not and cannot understand the how and why of them. (Jackson, 1992, 475)

It is his response to this worry that is of interest to us. Being unable to understand epiphenomenal qualia, he argues, is no objection to the claim that epiphenomenal qualia exist. This sort of thinking presupposes an excessively optimistic view of our epistemic capacities, capacities that, he claims, are circumscribed by our evolutionary past. He writes:

> We are the products of Evolution. We understand and sense what we need to understand and sense in order to survive. Epiphenomenal qualia are totally irrelevant to survival. At no stage of our evolution did natural selection favour those who could make sense of how they are caused and the laws governing them, or in fact why they exist at all. And that is why we can't. (Jackson, 1992, 475)

In short, the conclusion is that we cannot understand the qualitative aspect of consciousness because we were not designed by evolution to do so. Why should we have been? If Jackson is right, qualia have no physical effects, so how could there be an evolutionary reason for equipping us with the cognitive resources required to understand them?

Although it is worth noting that Jackson's argument, if successful, might leave much of consciousness untouched and therefore open to understanding -- it is not all of consciousness that is beyond our ken, on a certain reading of Jackson, just its qualitative properties -- qualia as he has characterized them are nevertheless a large part of consciousness. Are we forced to admit that qualia cannot be understood? Let us consider this argument in more detail.

More carefully, the line of thinking seems to be this. We are members of a species that has the characteristics it has because of the workings of natural selection. Let us understand his claim as innocuously as possible, perhaps along these lines: we are the result of a continuing process whereby environmental pressures selectively retain individuals in virtue of certain genetically based features that confer some survival and reproductive advantage over other individuals. There is much more to the story than this, obviously, but a sketch will do for now.

If something like this is true, then our perceptual and conceptual systems have the characteristics they have in

virtue of a selective process constrained by the need to survive and reproduce. We are, Jackson urges, made to understand only what we need to understand in order to survive. Understanding qualia has nothing to do with such needs. So we are not made to understand them. That is why we cannot understand them. The argument seems to give us evolutionary ground for thinking that at least some aspects of conscious experience are beyond our understanding.

It is worth dealing with a certain ambiguity in the passage before considering the argument in more detail. He claims that qualia are totally irrelevant to survival. Does this mean that understanding qualia is irrelevant to survival or that qualitative states themselves are irrelevant to survival? Jackson would probably endorse both claims, but the argument is clearly for the former. He is saying that, because qualia are irrelevant to survival (they have no physical effects, on his view, so how could they be relevant?), and our powers of understanding are geared towards that which is relevant to survival, we are not designed to understand qualia. With this in mind, let us look at the conclusion of the argument and the argument itself in more detail.

Two questions of interpretation suggest themselves. What does the claim that qualia cannot be understood amount to for Jackson? How are we to understand the argument for that claim? Let us consider each of these questions in turn.

Interpretation of the conclusion

First of all, then, what is it that cannot be understood? The line quoted above is suggestive: 'At no stage of our evolution did natural selection favour those who could make sense of how they [qualia] are caused and the laws governing them, or in fact why they exist at all.' (Jackson, 1992, 475) Perhaps what Jackson means when he says we cannot understand qualia is that we cannot know (1) how they are caused; (2) the laws that govern them, and (3) why they exist.

By (1), how qualia are caused, he must mean something more than the obvious physical causes of phenomenal experiences. Jackson admits that smelling a rose is the cause of the smell of a rose, that getting hurt is the cause of the hurtfulness associated with pains, that tasting a lemon is the cause of the taste of a lemon, and so on. He knows that qualia are caused and what, in general, they are caused by; his claim is not that they are causeless, but that they have no physical effects.

In addition, he must mean something more than the detailed causal story we can tell that begins with smelling a rose or getting hurt (or whatever impingements on our nerve endings are in question) and proceeds through our sensory systems into characteristic patterns of neural activity. He must also mean something more than the causal connection that he admits exists between those patterns of neural activity and

qualia -- certain happenings in the brain cause qualia, he says. It is not clear what is left for him to deny. He admits that we know that something in the world causes the stimulation of our sensory apparatus, that our sensory apparatus then causes certain neural events, and that those neural events cause qualia. So, if we know all that, do we not know how they are caused?

Perhaps, when he says that we cannot know how they are caused, he means that we cannot know a part of the causal chain, a middle part that we know exists but nevertheless cannot grasp. In particular, perhaps we cannot know the steps mediating happenings in the brain and phenomenal experience, what causes the particular quale to arise as a result of brain events and the rest of the causal story that we can tell.

If this is the case, then he thinks qualia are caused by physical events, but that there is a part of that chain that we cannot know, a middle part, that we know exists, but nevertheless cannot fathom. We know the beginning. We know the end. We know a great number of the intermediate steps. If Jackson admits all this but still holds that we cannot know how they are caused, then the claim might be that we cannot know one or more of the intervening steps. It is important to emphasizes the oddness of that assertion, and how difficult it is to think of examples of other causal chains that we know about but have parts that are hidden away, in principle. If Jackson is right, and we understand what we need to understand

in order to survive (more on this in a moment), one wonders what is it about the parts of the causal chain that we do know that is relevant to survival, and what is it about the parts of the causal chain that we cannot know that is irrelevant to survival. No obvious answer suggests itself.

Similar questions arise in connection with (2), the claim that we cannot make sense of the laws that govern qualia, if Jackson is referring to causal laws. If the above story is right, we know quite a bit about the laws associated with qualia: certain sorts of neural stimulation cause pain; cones stimulated by light waves within a particular band cause the quale associated with seeing blue, and so on. We exploit knowledge of such laws when we build television screens, anaesthetize patients, produce artificially flavoured crisps, and so on.

Again, perhaps Jackson is talking about a hidden part of the causal chain, and the laws specifying connections that range over hidden parts of the causal chain. Whatever reservations one might have about (1) seem to resurface with (2) as well.

Finally, Jackson claims that we cannot understand qualia in so far as (3), we cannot know why they exist. Traditionally, why questions are interpreted in two ways: causally or teleologically. If Jackson means the former, then understanding why qualia exist is being able to give a causal account of them -- being able to say what causal sequence

brings them about. This line of thinking throws us back to the worries raised with respect to (1) again.

If Jackson is saying anything more than (1) here, then a teleological interpretation might be applicable: understanding why qualia exist is being able to say what purpose or function qualia serve in our cognitive economy. So perhaps Jackson is claiming that we cannot understand qualia insofar as we cannot say what purpose they serve.

Jackson himself provides a response to this sort of worry. He reasonably assumes that having qualitative properties is something organisms have evolved over time -- we have qualia, he argues, and the earliest forms of life did not. If this is true, then certain things follow about the purpose of qualia, on his view. Jackson maintains that 'we should expect any evolved characteristic to be either conducive to survival or a by-product of one that is so conducive.' (Jackson, 1992, 474) So either qualia are conducive to survival or they are by-products of something that is conducive to survival.

Why are those two possibilities consistent with the claim that qualia have evolved? Recall Jackson's example: polar bears have thick, warm coats, and having a thick, warm coat affords polar bears considerable survival advantage. A by-product of having a thick, warm coat is having a heavy coat. Having a heavy coat is not conducive to survival, because it slows polar bears down. The advantages of having a thick,

warm coat outweigh the disadvantages of having a heavy coat. So it is consistent with what Jackson believes about evolution for polar bears to have thick, warm but heavy coats. Traits are either advantageous or the by-product of traits that are advantageous.

Jackson holds 'that qualia fall into the latter category. They are a by-product of certain brain processes that are highly conducive to survival.' (Jackson, 1992, 474) So Jackson does have a teleological understanding of qualia and an answer to the question, why do qualia exist. According to him, qualia exist because they are natural concomitants of neural events that confer some survival advantage.

So it is difficult to see exactly what Jackson thinks we cannot know, given the three points he mentions in this connection. As we have just observed with respect to (3), Jackson does understand why qualia exist insofar as that notion is to be interpreted teleologically. He maintains that qualia are the by-products of certain brain processes that are conducive to survival. As we have seen with respect to (1) and (2), he admits that we understand much of the causal story and the laws that attend it. So it is not clear exactly where our ignorance lies. He commits himself to the view that the stimulation of sensory systems causes brain events which cause qualia. What remains? We are left with the possibility that we know that one event causes another through a series of steps, but that at least one of those steps is hidden from us

in principle. Such a view is obscure at best.

Interpretation of the argument

Let us now turn to Jackson's argument for this conclusion. He writes:

> We are the products of Evolution. We understand and sense what we need to understand and sense in order to survive. Epiphenomenal qualia are totally irrelevant to survival. At no stage of our evolution did natural selection favour those who could make sense of how they are caused and the laws governing them, or in fact why they exist at all. And that is why we can't. (Jackson, 1992, 475)

The claim that we are the products of evolution is innocuous. Further, we can, for the sake of argument, grant Jackson's claim that qualia are irrelevant to survival. Though we have seen that a number of problems attend the epiphenomenalist view, that is not where the quarrel is, at the moment. What is interesting about Jackson's view is not his commitment to epiphenomenalism, but his claim that we cannot understand a certain aspect of consciousness.

The proposition that natural selection never favoured those of our forbears who could understand qualia just in so far as they could understand qualia seems to follow from the

claim that qualia are irrelevant to survival, in conjunction with certain bare presuppositions about natural selection.

Much seems to hang on the claim that 'we understand and sense what we need to understand and sense in order to survive'. This is the crux of things for Jackson and for others (for example, McGinn, 1993) who maintain that evolution gives us reason for skepticism about understanding consciousness.

What exactly does Jackson's crucial claim mean? Although it is true that we, in some sense, must understand what we need to understand in order to survive -- we do survive, after all -- one wants to say that this cannot be all there is to human understanding. It seems clear that we do understand a great deal that is plainly irrelevant to survival: ballet, the ontological argument, particle physics, haiku, Hubble's law, football, fermentation and so on.

The reason for this tension is that there are two ways to interpret the claim that we understand and sense what we need to understand and sense in order to survive. The claim might be taken in either of the following two ways:

(A) We understand and sense *at least* what we need to understand and sense in order to survive.

(B) We understand and sense *only* what we need to understand and sense in order to survive.

(A) is a claim about the bare bones of our conceptual and perceptual systems. It amounts to saying that, at minimum, we understand and sense what we need to understand and sense in order to survive -- and we might understand and sense more besides. Given certain assumptions about evolution, that claim is plainly true, and this is probably what gives the argument an air of plausibility. Our species survives, so of course we understand and sense what we need to understand and sense in order to survive. However, on this interpretation, the conclusion does not follow, because the argument does not rule out the possibility of our understanding more than what is relevant to survival. So we must not interpret Jackson along the lines of (A).

We seem forced to understand him in terms of (B), a claim about the extent of our conceptual and perceptual systems, about the very limits of our understanding: we understand and sense nothing more than what we need to understand and sense in order to survive. If that were true, and understanding qualia were irrelevant to survival, then the sceptical conclusion would follow. However, now the premise seems false. Think again of ballet, the ontological argument, particle physics, haiku, Hubble's law, football, fermentation, and the very many other things we understand that seem irrelevant to survival.

4.3. The Adaptationist Fallacy

Given this distinction, we are in a position to draw conclusions about Jackson's argument. The fundamental difficulty with the argument is that it seems plain that we understand much more than we need to understand in order to survive. Is there reason to think that the premise actually is false, is there something more than the fact that it seems to us that we understand much that is irrelevant to survival?

Jackson himself provides a reason for thinking that (B) actually is false. Recall his reply to the second objection considered in part two of this chapter and the discussion of the teleological interpretation of the claim that we cannot understand why qualia exist. Jackson maintains that:

> ...all we can extract from Darwin's theory is that we should expect any evolved characteristic to be either conducive to survival or a by-product of one that is so conducive. (Jackson, 1992, 474)

If Jackson is committed to this understanding of our evolved properties, then he must hold that our cognitive faculties are either conducive to survival or a by-product of some characteristic that is conducive to survival.

If that is true, then it is possible that our understanding is not constrained just by survival needs. Our understanding might be the way it is not because selection pressures constrain its application, but because it or parts

of it are a side-effect of some other cognitive capacity that has survival value. The capacities required to understand qualia might not be relevant to survival, but Jackson has not ruled out the possibility that such abilities are a consequence of cognitive faculties that are conducive to survival.

The possibility is certainly a live one. Given the very many things that we seem to understand that are not obviously relevant to survival, one might suspect that a significant portion of our cognitive endowment is not directly engaged in the bare business of existing. Until Jackson produces an argument that rules out this possibility, the conclusion that qualia cannot be understood does not appear to follow.

So on either interpretation of the premise, the argument seems to fail.

General difficulties with arguments from evolution

This problem might be cast differently, abstracted away from Jackson's version, such that it applies to sceptical arguments from evolution generally. Either the skeptic is claiming that all human traits relevant to understanding consciousness are selected traits or that some are not. If it is the latter, then the possibility exists that some relevant properties are not constrained by biological need, that our capacities might overreach the bare necessities of survival and reproduction. In this case, there is every chance that

our understanding can go beyond what is relevant to survival.

If, however, it is maintained that all relevant traits are selected traits, then the skeptic is committed to a kind of dubious adaptationism, and there are good reasons for thinking such a view is simply false.

Darwin himself thinks as much, as the following passage indicates:

> [I]n the first edition of this work [Origin of Species] I placed in a most conspicuous position -- namely, at the close of the introduction -- the following words: "I am convinced that natural selection has been the main but not the exclusive means of modification." This has been to no avail. Great is the power is misrepresentation. (Darwin,[1859], 1972, 115)

Assertions in a similar spirit to the all or nothing interpretation of Jackson's claim [version (B)] presuppose that all traits exist in virtue of selection pressure, that all traits confer some survival value. This is no part of either classical or contemporary evolutionary theory.

Darwin knew that there was more to modification than selection pressure, that some mechanism underlies the inheritance of traits, but he did not understand it. Geneticists more or less do. It is now clear that evolution

sometimes results from mechanisms other than natural selection, and that traits arise and persist that are not fixed by basic survival needs.

There are many well known processes of non-adaptive evolution: germ line mutation, gene flow, the cessation of gene flow, random genetic drift, mutation pressure, Kimura's neutral mutation, and the founder effect, to name a few. Take the founder effect, for example. Suppose a few individuals (maybe even a single pregnant female, as might be the case with the single forbear from which it is believed all modern cheetahs are descended) are accidentally separated off from the rest of the population. The new, founder genome will probably not be representative of the parent population, and inbreeding early in the new population's history will lead to increased homozygosity. Gene frequencies will probably differ radically between the two populations, as a result of happenstance, not natural selection.

Besides such processes, phylogenetic inertia ensures that evolution does not produce optimally designed organisms. Evolved creatures are historical objects, as Stephen Jay Gould points out in many publications, and some of their structures eventually become functionally useless or get pressed into service in novel ways. An organ may no longer serve a function at all, but if it makes no difference in terms of the creature's fitness, evolution might ignore it.

Of course, if the modification is maladaptive then

natural selection is likely to weed it out of the gene pool, but a great many structures persist in selective limbo. Think of the human appendix, a legacy that was once probably useful, but now serves no obvious function at all. It is also true that some structures fulfill functions that they were not originally designed for. Organisms are sometimes jury-rigged with ill-fitting parts marginally accomplishing tasks for which they were not originally geared.

Evolved structures sometimes do considerably more than they were selected to do. Fingers type, for example. This is precisely what some philosophers think is the case with our conceptual abilities. Dennett (1991), for example, argues that the mind can best be understood as a virtual machine implemented in a brain that was not designed for such activities. It is not a particularly helpful metaphor, but, in broadest outline, the idea is that the mind is like a program running on hardware that was not designed for the task. The hardware was designed by natural selection just to understand what is necessary for survival, but our minds now do much more than that.

Using Jackson's language, a by-product of this bare capacity turned out to be a kind of plasticity that enabled our forebears to rely less on hardwired solutions to the problems presented by the world and more on the ability to deal with problems as they arose, even think about potential problems before they become problems.

Think of the Sphex wasp (Hofstader, 1989) to appreciate the survival value accompanying plastic as opposed to hardwired solutions to problems. The creature stings its prey and drags it back to the nest. The wasp then drops it outside and looks inside the nest, apparently checking for predators. If all is clear, the wasp drags its prey inside and lays eggs. Researchers have discovered that the wasp will repeat this procedure indefinitely if the prey is moved away from the nest while the wasp is inside. Its apparently intelligent behaviour is actually a tropistic reflex, and the wasp will literally starve to death moving its prey back and checking its nest, if the prey is continually moved while the wasp is inside.

Flexible cognition, to follow the line of thinking through, brings with it something more than the capacity to think about problems relevant to survival -- a really flexible mind can think about things that make no difference at all in evolutionary terms. Once thoughts like these enter the picture, it becomes possible to understand more than is needed in order to survive. Once groups of flexible minds begin sharing information, writing things down, creating things, developing new strategies for formulating and dealing with problems, developing a culture -- once all of these things start happening, the understanding is no longer circumscribed by the need to survive and reproduce. Its constraints have shifted.

I am not arguing that this picture of our conceptual faculties is exactly right, only that it or something like it seems possible or even likely -- that our capacity to understand (like so many of our evolved capacities) might not be the way it is solely in virtue of selection pressures. Perhaps plasticity of mind brings with it as a side-effect the ability to compose music, write poems, demonstrate theorems, appreciate sunsets, and philosophize. As long as the side-effect story of our cognitive abilities is a viable one, the sceptical conclusion can be resisted.

There is an even stronger line one might take against the view that the capacities needed to understand qualia have no survival advantage and are, therefore, no part of our cognitive endowment. It might be claimed that the capacities needed to answer certain philosophical questions are, in fact, quite useful from an evolutionary standpoint. The ability to make clear distinctions, identify hidden connections, follow an inference through and the rest of the abilities touted in first year philosophy course handbooks are all abilities that would bring some survival advantage. Maybe the same capacities needed to understand qualia are just those needed to think through certain problems associated with surviving.

Conclusions

Before we consider the general conclusion of this chapter, it might be worth asking whether or not the problems

associated with the argument from evolution have any consequences for the original knowledge argument itself. It might be, as suggested at the outset, that the argument from evolution is not a straightforward argument on its own for the claim that consciousness cannot be understood, but an explanation for why it cannot be understood in a certain sense, given the conclusion of the knowledge argument. In other words and contrary to the interpretive line of this chapter, it might be that Jackson is using Darwin to explain why consciousness cannot be understood, rather than using Darwin to show that consciousness cannot be understood -- the knowledge argument, not Darwinian considerations, accomplishes the latter task.

If this is the right way to interpret Jackson, then the failure of the argument from evolution might have consequences for the knowledge argument. Recall that Jackson brings up the argument from evolution in the first place in response to the worry that qualia 'do nothing, they explain nothing, they serve merely to soothe the intuitions of dualists...in short we cannot understand the how and why of them' (Jackson, 1992, 475). Without the argument from evolution to explain why qualia are inexplicable, a certain kind of pressure is placed on the knowledge argument and its conclusion that qualia are epiphenomenal -- epiphenomenal qualia seem to exist only to soothe the intuitions of dualists.

If Jackson is left without an explanation of our failure

to understand consciousness, then either there is some other explanation, or consciousness can be understood. Until Jackson provides us with another explanation, we might suppose that consciousness can be understood, in which case there must be something wrong with the knowledge argument. As we have seen, there are a number of objections to it. The failure of the argument from evolution -- Jackson's failure to provide an explanation of why consciousness cannot be understood, on this reading -- might strengthen the force of those objections, if only a little.

At any rate, we have considered Jackson's epiphenomenalist position, his argument from evolution, and his conclusion in some detail. The argument from evolution faces a number of problems. The conclusion, first of all, that qualia cannot be understood, proves on inspection to amount to something fairly puzzling: the claim that some part of a causal chain is hidden from us, in principle. How such a thing could be remains obscure.

The argument itself appears to be either invalid (with version A of the second premise) or fallacious (with version B). This is not just a problem for Jackson. Similar arguments from evolution that presuppose that the understanding is constrained just by the need to survive and reproduce are also guilty of an adaptationist assumption -- the proposition that all traits exist exclusively in virtue of the need to survive and reproduce. The story is probably much

more complex than that, as we have seen. Therefore, arguments from evolution to conclusions about the prospects of understanding consciousness are at this stage unconvincing.

5. Cognitive Closure and the Mind-Body Problem

Are there any other reasons that might be thought to support the claim that consciousness cannot be understood? If we reflect on the prospects of coming to understand some phenomenon or other, we might be swayed by the following two thoughts.

1. Our minds are limited in various ways. Memory and perception operate under constraints that we experience daily, not the least of which are the simple limits of the channel capacity of our neural pathways and the thresholds of our discriminatory faculties. Perhaps more interesting than this, for our purposes, is the possibility that our minds have certain conceptual limitations. That is to say that, given the structure of our minds, we have access to a limited store of concepts. Given the fact that our minds are limited in the ways that they are, it seems reasonable to suppose that some concepts are simply beyond us.

2. Reality is what it is regardless of our conceptual discriminations. How we think about the world has no bearing on the way that the world is. We might gain cognitive access to different aspects of the world when we think with different concepts,

> but the aspects themselves do not change when we apply concepts to them. Further, it seems obvious that some parts of the world only come into view when we think of them with the right concepts. The point is that those parts of reality are there whether we manage to think with the requisite concepts or not.

If it is true that reality is the way it is regardless of how we think it to be, and we have only a limited store of concepts to bring to bear on the world, then it seems at least possible that some part or parts of the world are beyond the limits of our understanding. There might be an aspect of reality that we cannot get our conceptual hooks into, because we have the wrong hooks for the job. That is to say that there might be aspects of reality thinkable only with concepts that the structure of our minds precludes us from having.

This suggests to some that there are questions we can ask but never answer to our satisfaction because the answers themselves require a conceptual innovation that is beyond our cognitive power. We might have the concepts needed to identify a problem area and so ask a meaningful question about it, but we might nevertheless lack the capacity to form the concept or concepts required for the answer. Reality might yield to our capacity to wonder while outstripping our ability to understand.

If this line of thinking is correct, then for any phenomenon we encounter, there is a sense in which we might never be able to form the concepts needed to understand it. Of course the conclusion that a given thing cannot be understood does not follow from this general possibility alone. What is needed is some reason for thinking that we actually are in the unfortunate epistemic situation of pondering a problem with a solution outside the boundaries of our conceptual space. Since our interest is consciousness, our question is this: are there any good reasons for thinking that the answers to the questions, what is consciousness and what is its relation to the body, require conceptual innovation beyond our cognitive capacities?

McGinn argues that there are good reasons, that the answers to these questions are beyond our conceptual powers. In this chapter, we will consider his main argument for the claim that we cannot solve what he calls 'the problem of consciousness', reserving a consideration of his argument for the claim that we cannot understand what consciousness is for the next chapter.

This chapter has three parts. In part one I will recapitulate McGinn's argument for cognitive closure. A certain line of inquiry is suggested by what appear to be several different possible closure candidates surfacing in the premises of the argument, and part two of this chapter is an attempt to identify and give some account of them. Part three

is an evaluation of the argument itself, based on the closure candidates identified. In the end, I hope to show that nothing McGinn says should lead us to the conclusion that the mind-body problem cannot be solved, on any of the possible interpretations of the problem considered.

5.1 The Argument for Cognitive Closure

Preliminary considerations

If we are to have a clear understanding of McGinn's main argument for the insolubility of the mind-body problem, we must distinguish it from a tangential argument that surfaces from time to time in McGinn's writings, an argument all too common in other discussions about the prospects of understanding consciousness. It is sometimes claimed that something about the prospects for understanding consciousness follows from the fact that we have, thus far, failed to understand it. McGinn writes,

> Longstanding historical failure is suggestive, but scarcely conclusive....I think that our deep bafflement about the problem, amounting to a vertiginous sense of ultimate mystery, which resists even articulate formulation, should at least encourage us to explore the idea that there is something terminal about our perplexity. (McGinn, 1994, 7)

Clearly, the conclusion that we cannot ever understand consciousness does not follow from the claim that we do not understand consciousness now, or even that we find consciousness difficult to understand. To think that it does is an obvious example of an argument from ignorance. The reason arguing from ignorance is fallacious is that nothing interesting about the nature of an *explanandum* or the prospects for understanding it follows from our current epistemic status -- even if we are currently absolutely baffled by the thing. We cannot infer, from our present failure to understand, that consciousness is different in kind from everything else that we do understand, or that consciousness will forever elude us. The only thing that follows from our longstanding failure to solve the mind-body problem is that if there is a solution, it eludes us, not that it will forever elude us.

Although McGinn adds the disclaimer that historical failure is scarcely conclusive, he does nevertheless maintain that it is suggestive and that our 'deep bafflement' should encourage us to explore the possibility that our perplexity is terminal. Nevertheless, our failure and our bafflement are not suggestive or encouraging in the sense McGinn seems to intend at all. Unless we are persuaded by arguments from ignorance, nothing about our future epistemic prospects follows from our present ignorance.

It might be claimed that the disclaimers point to a

different reading of the passage: McGinn is not arguing for terminal mystery here, he is only claiming that the place to look for terminal mystery is where we presently fail in our efforts to answer questions. If that is so, then the passage is innocuous, but trivial. Of course the place to look for terminal mystery is in the long catalogue of items we do not currently understand -- obviously, it would be ridiculous to consider phenomena we already understand as candidates for things that cannot be understood. Nevertheless, we need something more than just ignorance or historical failure to go on, so let us consider McGinn's main argument for the claim that the mind-body problem cannot be solved.

The thinking underlying the argument is clearly rooted in a commitment to naturalism. For McGinn, 'Naturalism in the philosophy of mind is the thesis that every property of mind can be explained in broadly physical terms. Nothing mental is physically mysterious'. (McGinn, 1991, 23) Our interpretation of McGinn's claim is hampered somewhat by the fact that 'naturalism' and 'physical' are understood in a variety of ways, and the nature of McGinn's commitments are not clear at the outset.

Many understand naturalism negatively, as a position consisting in the denial of the existence of the supernatural. The supernatural is then identified by example -- in this connection, souls, spirits, mindstuff and the like. This is clearly part of McGinn's outlook, for he claims to be

'resolutely shunning the supernatural' (McGinn, 1994. 6) when speculating about the mind-body relation. Given only this, however, it is not clear what the natural is, or indeed, what it is about the supernatural that renders it suspect.

When formulated positively, naturalism often entails allegiance to physicalism, and, given his language, this too seems to be part of what McGinn means by 'naturalism'. However, physicalism, no less than naturalism, is understood in a variety of ways, and McGinn does not specify his conception of physicalism at the outset. Ultimately, the difficulty lies in specifying what it is for a thing to be physical. Minimally, the physical might be spelled out in terms of paradigm cases: objects that figure into scientific theories, such as electrons, elements, spleens, and fish. As we saw in the last chapter, this alone, though helpful at least to focus our attention a little, leaves us with no clear idea of what it is about these examples that makes them physical. It is sometimes claimed that such objects are physical because they all occupy space and time and can stand in causal relations with one another. So as to beg as few questions as possible, let us leave the physical construed in this way: the physical consists in space-time objects that can, in principle, stand in causal relations with one another.

McGinn is also silent about what an explanation is, so it is fair to wonder what the epistemic component of physicalism comes to for him. He claims that every property of mind can

be explained in physical terms. Again, so as to beg as few questions as possible, we shall leave explanation largely unspecified, hoping to fill in the gaps when necessary as McGinn's argument takes shape. For now, when McGinn claims that mental properties can be explained in broadly physical terms, let us suppose that he means that no entities other than the sorts of space-time objects described above need be invoked in explanations (however characterized) of mental properties.

There is something slightly worrying about this, though, because McGinn claims not that every property of mind can be explained in physical terms, but that every property of mind can be explained in *broadly* physical terms. What a broadly physical explanation is, as opposed to a physical explanation, remains obscure. Given McGinn's flat denial of the existence of anything supernatural (and his endorsement of the principle of homogeneity, which we will come to in a moment), he must not mean that such explanations invoke mostly physical objects, and a few souls as well. Perhaps he means to take account of the possibility of such things as functional and computational properties, supposing that, though they are physically instantiated, it is best to think of them as nonphysical in some sense. At any rate, nothing much turns on the word 'broadly' in his argument, so for our purposes, let us leave our understanding of explanation outlined above unmodified.

Let these be the default conceptions of naturalism, physicalism and explanation operative in our understanding of McGinn's thinking, with the proviso that they might have to be amended as the argument progresses.

An account of the argument for closure

The argument begins with the claim that 'there exists some property, P, instantiated by the brain, in virtue of which the brain is the basis of consciousness.' (McGinn, 1991, 6) Unless we are willing to countenance miracles or go eliminativist with respect to consciousness, McGinn claims, there must be a natural solution to the mind-body problem. 'P', for McGinn, is the key to it.

We know that supernatural answers to our questions convey no real understanding, and we know that consciousness exists. So if we are going to have a solution to the mind-body problem at all, McGinn argues, it must be something recognizably like the natural solutions to other problems we are familiar with in the history of science, solutions that invoke relevant space-time objects and characterize their interactions in a certain way. We observe correlations between consciousness and brain states, and unless we are to assume that such correlations are just brute facts about the way the world works, there must be some natural explanation that accounts for them, or so McGinn maintains. Just as we reject the claim that life arises in virtue of divine intervention and search

for natural explanations for its emergence, he maintains, we reject the claim that consciousness has some supernatural basis. There is something about the brain -- some property no more metaphysically extravagant than the many neural, chemical, functional properties and the like that we have already identified -- that makes the brain the basis of consciousness. If we could get into the right epistemic relation with the requisite natural explanatory property, we could have a solution to the mind-body problem. The question is, can we get into that relation?

In the next step of the argument, he delineates what he takes to be our only routes to the desired epistemic relation, the only ways in which we might identify the crucial explanatory property. He maintains that 'there seem to be two possible avenues open to us in our aspiration to identify P: we could try to get P by investigating consciousness directly; or we could look to the study of the brain for P'. (McGinn, 1991, 7) The question then becomes, can these two faculties deliver P?

In the third step of the argument, he takes up what he characterizes as direct investigation of conscious states, introspection. In virtue of this faculty, he maintains, we have unmediated acquaintance with at least some of the properties of consciousness, but not with P. In McGinn's words, through introspection 'we have direct cognitive access to one term of the mind-body relation, but we do not have such

access to the nature of the link'. (McGinn, 1991, 8) The idea seems to be that since P is the link between the mind and the brain, and since an introspective examination of the mind only puts us in touch with one of the things linked, introspection cannot deliver the link itself.

If we cannot introspect the link, can we not arrive at a conception of it via some chain of inferences that begins with what we can introspect? He asserts that the effort to distill P by analysing the concepts of consciousness we already possess in virtue of being conscious just 'does not seem feasible'. (McGinn, 1991, 8) Trying to solve the mind-body problem by ruminating on phenomenal concepts seems as fruitful a strategy to McGinn as trying to understand how living things arise from inanimate matter by reflecting on the concept 'life'. The third step of the argument, then, is that neither pure introspection, nor reflection on the concepts that we deploy in the first person ascription of conscious states can deliver P.

Next, McGinn considers the prospects for the empirical study of the brain in our efforts to identify P. He begins by arguing that P is perceptually closed. What he means by 'perceptual closure' is fairly straightforward. The human perceptual apparatus is sensitive to only a limited set of properties. Our eyes are tuned to a narrow band of the electromagnetic spectrum, for example, and we are, therefore, perceptually closed to all wavelengths other than those which

comprise visible light. So when McGinn claims that we are perceptually closed to P, he means that we cannot discover P via the direct action of our sense organs alone. Just as we cannot see microwaves spilling out of a faulty microwave oven, we cannot detect P by examining the brain -- by simply looking, touching, and so on.

His reason for this claim begins with an appeal to the imagination. He says that we seem to be incapable of imagining a perceptual property instantiated in the brain that would assuage the feeling of mystery attending the mind-body relation. 'It is like trying to conceive of a perceptible property of a rock that would render it perspicuous that the rock was conscious'. (McGinn, 1991, 11) The idea seems to be that for any proposed property we might detect in the brain, we could never imagine such a thing rendering consciousness intelligible. The reason for this is, he argues, that our perceptual faculties necessarily represent properties in space, but consciousness does not seem to be made up of those sorts of properties. Thus, they are the wrong kind of things for the explanatory job. He writes:

> Consciousness does not seem made up out of smaller spatial processes; yet perception of the brain seems limited to revealing such processes. The senses are responsive to certain *kinds* of properties -- those that are essentially bound up with space -- but

these properties are of the wrong sort (the wrong *category*) to constitute P. (McGinn, 1991, 12)

In short, McGinn is saying that since consciousness is not built up of spatial properties, a faculty that only delivers spatial properties cannot give us the property that we seek.

However, even if we are perceptually closed to P, it might still be possible to form a conception of P -- the fact, if it is a fact, that we cannot sense P does not entail the proposition that we cannot arrive at a conception of P by some other means. After all many theoretical objects that we have a good cognitive grasp of cannot be directly perceived: consider subatomic particles, the Big Bang, radio waves and the like. Since we regularly make inferences from what we observe to what we cannot, McGinn needs to go further and argue that we cannot arrive at P by hypothesizing from the observable to the unobservable.

He does so by claiming that the introduction of theoretical concepts based on observation is constrained by what he calls the principle of homogeneity. He explains the principle by citing Nagel: '"it will never be legitimate to infer, as a theoretical explanation of physical phenomena alone, a property that includes or implies the consciousness of its subject"'. (McGinn, 1991, 13) As McGinn puts it: 'purely physical data will never take us outside the realm of the physical' (McGinn, 1991, 13), forcing us to introduce

another kind of concept, *i.e.* the concept of consciousness. Physical events have purely physical explanations, and 'since we do not need consciousness to explain those data, we do not need the property that explains consciousness'. (McGinn, 1991, 13) So inferences from the observation of the physical features of brains can never give us P.

All of this leads McGinn to the fourth major claim of the argument: perception and perception based inferences cannot deliver P. Since he has also claimed that introspection and perception are the two avenues open to us in our efforts to identify P, and introspection cannot deliver P either, he claims to secure the basis of his skeptical position. The mind-body problem cannot be solved.

Noumenal naturalism

What sort of overall view emerges from this argument? Consider McGinn's own account of his position. He maintains that there are two traditional ways of dealing with the mind-body problem. The first relies on the supernatural -- it invokes spirits, souls, or a god of the gaps in the hope of somehow making sense of the relation between consciousness and the body. McGinn rejects such solutions because they not only replace one mystery with another, but they do not square with what we already know about the physical world. Consciousness, he says, 'must be a natural phenomenon, naturally arising from certain organizations of matter.' (McGinn, 1991, 6)

So one expects McGinn to advocate the second kind of position: what he calls constructive accounts of consciousness, which explain how consciousness arises by identifying some natural property of the brain. However, the position he adopts is 'naturalistic but not constructive: I do not believe we can ever specify what it is about the brain that is responsible for consciousness, but I am sure that whatever it is it is not inherently miraculous.' (McGinn, 1991, 2)

The problem is not with the way the world is, but with the way our minds work. Though there is some natural property of brains in virtue of which brains give rise to consciousness, we are, as he says, cognitively closed to it. According to his definition, 'a type of mind M is cognitively closed with respect to a property P (or theory T) if and only if the concept-forming procedures at M's disposal cannot extend to a grasp of P (or an understanding of T).' (McGinn, 1991, 3)

In virtue of such things as selection pressures, there are different kinds of mind, and each one has different characteristic talents and endowments, as well as different deficiencies and weaknesses. These talents and deficiencies determine the extent to which a given creature has cognitive access to the world. Monkey minds, for example, are not up to the conceptual demands of atomic theory, and properties such as being an electron are cognitively closed off from their

apprehension. Everything that we explain in terms of atomic theory is relegated to the status of eternal mysteries for monkeys, in so far as atomic theory is beyond their concept forming capacities

As monkeys are to atomic theory, so we are to the solution to the mind-body problem, or so McGinn claims. Solving it requires a conceptual innovation precluded by the limits of the concept forming capacities at our disposal. There is nothing miraculous about the relation of the mind and brain, but the solution lies in our cognitive blind spot, and this makes the relation seem miraculous to us.

McGinn's claims about the prospects for understanding consciousness are part of a larger, metaphilosophical view, called transcendental or noumenal naturalism. It seems to have two central tenets, one ontological and the other epistemological. All properties are natural properties, hence the view is a species of naturalism; and some of those properties are inaccessible to the human mind, hence 'transcendental' and 'noumenal' are apt modifiers.

This general position is advanced as an hypothesis concerning the origins of philosophical perplexity. The explanation for what McGinn identifies as the special hardness of philosophy has nothing to do with the way the world is, but with the way our minds are. McGinn puts it this way:

...philosophical perplexities arise in us because of

> definite inherent limitations on our epistemic faculties, not because philosophical questions concern entities or facts that are intrinsically problematic or peculiar or dubious....Reality itself is everywhere flatly natural, but because of our cognitive limits we are unable to make good on this general ontological principle. Our epistemic architecture obstructs knowledge of the real nature of the objective world. (McGinn, 1993, 2)

McGinn holds that the (perfectly natural) answers to some questions are beyond the reach of human faculties; the answers transcend our comprehension. There is nothing miraculous about the phenomena that elude us; it is the cognitive deficits peculiar to the human mind that circumscribes our understanding. The precise nature of our cognitive limitations and the way in which McGinn believes our cognitive architecture sometimes obstructs an understanding of the world will be of concern in a moment. First, consider two analogous kinds of limitation that McGinn invokes in an effort to explain the notion of a cognitive limitation.

He sometimes compares cognitive limitations to perceptual limitations, a notion which, as we have seen, appears in his main argument for closure. Different species are equipped with different perceptual apparatuses and, thus, have different perceptual access to the world. Bees can see into

the ultraviolet end of the spectrum, and, given the structure of human eyes, we cannot -- ultraviolet light is beyond our perceptual sensitivity. There is nothing in the world that is inherently imperceptible -- rather, it is the representational capacities of perceivers that define the limits of the perceptible world for a given creature. So too, McGinn maintains, with cognitive capacities.

Just as there are different kinds of perceptual apparatuses, there are different kinds of mind. Just as a creature's perceptual apparatus facilitates access to certain properties in the world, but not others, a creature's cognitive apparatus facilitates a grasp of certain properties in the world, but not others.

McGinn sometimes compares cognitive limitations to motor limitations as well. Creatures with different kinds of mind move, as McGinn puts it, through different cognitive spaces, just as creatures with different kinds of motor apparatuses move through different physical spaces. Birds and fish are literally geared to move through certain physical spaces, but not others, and creatures with cognitive abilities of particular kinds are geared to move through certain cognitive spaces, but not others. Again, it is the structure of the mind, not the nature of the world, that is responsible for epistemic limitations.

McGinn's metaphilosophical claims concerning noumenal naturalism are peppered everywhere with disclaimers, making it

clear that what he is up to is speculation, not the generation of proofs. What he does make plain is a view of the mind as an evolved organ like any other we possess, which exists because of certain adaptive functions it performs. All such organs exhibit limitations of some sort: dependence on certain contexts, stress levels, fatigue patterns, output maximums and so on. All such organs perform better under certain conditions than others. The immediate question all of this raises is whether or not the human mind works well under what might be called philosophical conditions. In other words; if the human mind is an evolved organ like any other, are its limitations such that problems of interest to philosophy fall outside the domain of problems the mind is adaptively geared to solve?

What sort of problem is the human mind geared to solve? McGinn attempts to answer this question with what he calls the 'CALM conjecture', an acronym which stands for 'combinatorial atomism with lawlike mappings'. The hypothesis is that the human mind is geared towards deploying a certain pattern of thinking when trying to solve problems, the CALM pattern: primitive elements are understood as standing in specific relations to complexes of those elements. McGinn cites physics, linguistics, and mathematics as theoretical domains that exhibit the CALM character. In each of these areas, smaller elements (particles, words, lines) come together in specific ways to from complex wholes (material objects,

sentences, geometric figures).

Our success in these fields is parasitic upon our general cognitive proficiency with understanding spatial relationships, particularly relations between parts and wholes, a facility that somehow confers appreciable evolutionary advantage. Areas such as these in which the CALM method works are domains in which the human mind excels, generating characteristically deep and satisfying understandings. McGinn's suggestion is:

> :..it is our conformity to CALM modes of thought that stands in the way of our achieving the kind of [philosophical] understanding we seek. That is the way our reason makes things intelligible to us, but in [philosophical] cases, the method breaks down, thus producing intractable puzzlement....We apply the CALM mode willy-nilly to our problems, but instead of solving them it only deepens our sense of perplexity. (McGinn, 1993, 20)

The human mind, skewed by selection pressures to deal with matter in space, with relations between parts and wholes, has a certain adroitness with problems soluble in CALM terms. Philosophical problems seem to resist this method, and they seem all the more mysterious given the success we have with CALM in other domains. If this line of thinking is correct,

part of the mystery of consciousness is engendered by our failure to understand it in CALM terms -- we cannot see how the combination of small elements like neurons results in conscious experience. Human reason produces deep understanding by deploying the CALM methods in some domains, but the subject matter of philosophy does not seem amenable to CALM analysis. The thesis McGinn is suggesting is that the kind of cognitive capacities evolved by minds under the selection pressures ours have been under are geared to solve CALM problems, but not philosophical ones.

As we have seen in our consideration of Jackson, arguments from evolutionary considerations do not prove very much about our ability to solve philosophical problems, particularly those surrounding the nature of consciousness. McGinn's agenda, though, is different from Jackson's, and this makes mounting a critical response to his evolutionary reflections difficult. It is clear that he is not arguing from the claim that our minds are limited to understandings grounded in our evolved capacities to deal with physical objects to the conclusion that philosophical problems are insoluble. He is, so he claims, attempting only to render plausible (not prove) the general notion that philosophical problems are hard because of our epistemic limits, not the world -- the evolutionary explanation of our difficulties is extremely speculative, according to McGinn. So long as this is born in mind, McGinn is actually saying very little that is

contentious. He says, for example,

> My aim...is to try out a very general hypothesis. The attitude I intend to produce towards the hypothesis is mere respect; if the reader ends up believing it, that is his or her own business. In the nature of the case, indeed, it is a hypothesis which does not admit of the kind of demonstration we naturally demand for hypotheses of its general form. My claim will be that it may be true.... (McGinn, 1993, 2)

McGinn's metaphilosophical claims are, therefore, extremely modest: transcendental naturalism may be true. Part of the reason why it may be true is that our minds may be limited by selection pressures.

Things become interesting when McGinn actually applies his general metaphilosophical claims to particular philosophical problems, arguing for the substantial claim that satisfying answers to philosophical problems associated with consciousness, the self, meaning, free will, the *a priori*, and knowledge lie outside the cognitive space accessible to the human mind. It is while engaged in this enterprise, actually trying to show that such problems are insoluble, that McGinn hopes 'to sow the seeds of philosophical respect' (McGinn, 1993, 2) for transcendental naturalism. If respect for the

metaposition is dependent upon the success of its particular applications, perhaps the best way to evaluate transcendental naturalism is by engaging in a critique of its many applications. This kind of response to his metaphilosophy is beyond the scope of this work, but insofar as McGinn's skepticism about the prospects for understanding consciousness is an application of transcendental naturalism, a critique of the former helps undermine the latter. This chapter, then, might go some way towards undercutting respect for McGinn's metaphilosophical position as well, but this can only be a subsidiary aim of the present work.

We have before us an account of McGinn's argument for the claim that the mind-body problem cannot be solved and an understanding of the larger metaphilosophical project in which this argument is embedded. So we are in a position to begin working towards an interpretation and evaluation of his argument.

5.2 What Does McGinn Think We Cannot Know?

A certain line of enquiry is suggested by a kind of incongruence between the conclusion of the argument for closure and a number of claims in the argument itself. (A condensed version of the argument in this section appears in Garvey, 1997.) McGinn maintains, at different points throughout his argument, that what we cannot do is identify a certain brain property, identify the nature of the link

between consciousness and the brain, and solve the mind-body problem. These seem, at first blush, to be different failings: being unable to identify a brain property seems to be different from being unable to identify a link between brain properties and conscious ones. The inability to solve the mind-body problem might be different yet again.

Once suspicions are raised and close attention is paid to what it is that McGinn claims cannot be done or known, a number of possibly different closure candidates can be discerned. Consider this list of candidates, formulated mostly in McGinn's own words: 'some natural property of the brain (or body) which explains how consciousness can be elicited from it' (McGinn, 1991, 2); 'a conception of that natural property of the brain (or of consciousness) that accounts for the psychophysical link' (McGinn, 1991, 2-3); an understanding of the 'kind of causal nexus' (McGinn, 1991, 3) existing between mind and body; 'a grasp of P (or an understanding of T)' (McGinn, 1991, 3), where P is the relevant explanatory property and T the theory in which it occurs; 'the connection between consciousness and the brain' (McGinn, 1991, 5); 'the correct explanatory theory of the psychophysical nexus' (McGinn, 1991, 5); 'some property of brains that accounts naturalistically for consciousness' (McGinn, 1991, 5); 'some theory T referring to P which fully explains the dependence of conscious states on brain states' (McGinn, 1991, 6); 'know[ing] T and grasp[ing] the nature of

P' (McGinn, 1991, 7); 'identifying P' (McGinn, 1991, 7); 'the nature of the link' (McGinn, 1991, 8) between consciousness and the body; 'the concept P' (McGinn, 1991, 8); 'the solution to the problem of how specific forms of consciousness depend upon different kinds of physiological structure' (McGinn, 1991, 9); 'the mind-body problem' (McGinn, 1991, 15); 'a constructive solution to the mind-body problem' (McGinn, 1991, 16); 'a certain science' that explains the nature of the psychophysical connection (McGinn, 1991, 17); 'the psychophysical mechanism' (McGinn, 1991, 19); and finally, 'how P is related to the "ordinary" properties of the brain' (McGinn, 1991, 20).

No doubt there is some overlap here, but there seem to be different conceptions of what we cannot do or know operative in McGinn's thinking. So what does McGinn think we cannot know? Let us distill the list of possible closure candidates down to the following three:

(A) We are cognitively closed to identifying the brain property in virtue of which the brain is the basis of consciousness.

(B) We are cognitively closed to the nature of the relation holding between consciousness and the brain.

(C) We are cognitively closed to solving the mind-body problem, saying how it is that brains

generate consciousness.

The question, what does McGinn think we cannot know, might be taken in two ways, both of interest for our purposes. First, we might ask, what claim is McGinn actually trying to secure with his argument? Given the introductory material and the title of the book from which the argument under consideration is taken, it is clear that McGinn is trying above all to show that we cannot solve the mind-body problem, formulation (C), in other words. Determining just what this comes to is a difficult matter, and we will deal with it in a moment.

The second and larger question, the one that will occupy us in the third section of this chapter, is this: it is clear that McGinn wants to show that the solution to the mind-body problem is beyond us, but what does McGinn's argument actually show?

The list just considered suggests that several conceptions of cognitive limitation or inability might be operative in McGinn's thinking. Let us begin with the most obvious closure candidate, (C), the claim that we cannot solve the mind-body problem. Then we will consider the other possibilities.

The mind-body problem

There is a difficulty, perhaps endemic in contemporary

philosophy of mind, in supposing that the mind-body problem is a single problem. The phrase, 'the mind-body problem', in fact points to a number of interrelated questions that might be asked about mind and body. For example, one understanding of the mind-body problem, perhaps the traditional understanding, concerns the nature of the general relation between consciousness and the brain. One might ask, among other things, if the relation is causal, if the mental and physical are identical, or whether or not the mental supervenes on the physical. On a more modern reading, the mind-body problem is prefaced with the notion that we already know the nature of the relation between mind and brain: it is causal. The question is about the precise nature of the causal process in virtue of which physical events bring about mental ones. How -- that is, by what process or mechanism -- do brains give rise to consciousness? On yet another reading, the question is not about process, but possibility: how *could* something so unlike the physical -- a given conscious mental event like feeling dizzy, say -- be caused by something physical? The question is asking not for a process, but an explanation of how such a process could exist in the first place. On yet another reading, the problem is epistemic: how can we render coherent our beliefs about the natural world and our beliefs about ourselves insofar as we are conscious subjects.

These questions, and no doubt many others, might be taken

to be *the* mind-body problem. Perhaps certain strategies result in the deflation of these problems to only a few, or perhaps just one. Without such strategies on the table, 'the mind-body problem' is ambiguous. So when McGinn argues that the mind-body problem cannot be solved, more must be said about what he means. Let us consider first what McGinn means by 'consciousness' and then, against that background, the sense in which he claims that the mind-body problem cannot be solved.

The problematic aspect of consciousness

At no point in McGinn's argument for cognitive closure does he attempt to say precisely what consciousness is. This would come as no surprise if McGinn's claim were that consciousness is radically unknowable, that is, if the whole of its nature were beyond us, and, thus, no general conception were available at all. So far as I am aware, no currently active philosopher holds this position, perhaps for good reason. If consciousness were radically unknowable in this way, such that there was no chance of even having a concept of it, it would be difficult to see what meaning could attach to claims about consciousness -- including the claim, 'consciousness is unknowable'. If we know nothing about what the referring expression stands for, it is unclear how to make sense of sentences in which it occurs.

At any rate, this is not McGinn's position. There must

be at least some conception of consciousness at work in his arguments. After all, his conclusion seems to be that we cannot solve a certain problem about consciousness. We must, therefore, have some conception of the problem area, the thing or property with which we have a problem. McGinn makes some claims about the nature of consciousness: it is a natural phenomenon; it arises in virtue of some brain property, and so on. Further, he claims that we 'have direct cognitive access' (McGinn, 1991, 8) to consciousness via introspection. So it is clear that, on McGinn's view, consciousness is not radically unknowable; we have access to it and some conception of it. Nevertheless, McGinn does not attempt to articulate the conception he has in mind.

He prefaces his main argument for closure with some mention of 'technicolor phenomenology' (McGinn, 1991, 1), which suggests that raw feels -- the way that colours, sounds, smells and the like seem to us -- are somewhere in the conceptual foreground. Moreover, he characterizes his interest as concerning 'subjective awareness' (McGinn, 1991, 1) which indicates that consciousness, for him, is somehow bound up with the notions of point of view and attention. This is all we have to go on when McGinn begins his argument for closure.

Fortunately, in a later work, *Problems in Philosophy*, he does say something about the aspect of consciousness that he claims cannot be understood -- in his words, 'the property of

consciousness that eludes physical explanation'. (McGinn, 1993, 28) So although he does not provide a conception of consciousness, he does claim to deliver a conception of that part of consciousness that he claims cannot be understood. Before considering this aspect, it is worth emphasizing that in identifying the problem area in this way, McGinn reduces the scope of his skepticism to just one aspect of consciousness. Whether or not the alleged problem area is a fundamental part of consciousness or merely some peripheral aspect is never made explicit by McGinn. Further, the rest of consciousness, that which is not picked out by McGinn's identification of the problem area, clearly remains untouched by his sceptical arguments.

The aspect of consciousness that eludes us, according to McGinn, is encapsulated in Nagel's phrase, 'there is something it is like to be an X'.(Nagel, 1991, 422) As this conception figures prominently in McGinn's thinking and in the thoughts of other skeptics, it is worth pausing for a moment to consider Nagel's formulation in some detail.

Nagel's conception is articulated in the following well known passage:

> ...the fact that an organism has conscious experience *at all* means, basically, that there is something it is like to *be* that organism....[A]n organism has conscious mental states if and only if

> there is something that it is like to *be* that organism -- something it is like *for the organism*. (Nagel, 1991, 422)

The expressions 'something it is like' or 'what it is like' are notoriously difficult to interpret. One immediately wonders how the word 'like' is being used.

A start on an understanding of the word might be made by thinking in terms of how 'what it is like' is used in everyday expressions. When it is said, for example, 'I understand what your childhood was like', often what is meant is that, given the things that you have said about your childhood, I see that my own childhood is similar to yours, resembles yours, is, in this sense, like yours. I recollect my own childhood and come to believe that it resembles yours in certain ways, and on the basis of this I claim to know what your childhood was like. However, interpreting the phrase with the notion of resemblance is expressly ruled out by Nagel. He writes:

> ...the English expression 'what it is *like*' is misleading. It does not mean 'what (in our experience) it *resembles*, but rather 'how it is for the subject himself'. (Nagel, 1991, 428)

If 'like' is misleading, what remains is a conception of 'what it is like' that is not understood in terms of resemblance,

but how it is for the subject himself. According to Nagel, then, 'what it is like to be an X' means 'how it is for an X'. The latter expression seems at first as unilluminating as the former. The fact that Nagel is willing to claim that the latter version is less misleading presents us with an interesting point.

The word 'it' is common to both versions, and one quickly wonders to what 'it' refers. What is *what* like, or, how is *what* for the subject, in other words? Perhaps it will be claimed that this line of questioning is too literal -- like demanding the referent for 'sake' when something is done for your sake. Given the great weight placed on the expression by many philosophers, McGinn among them, and given the fact that the word 'it' appears in both the original formula and in Nagel's clarification, it seems at least fair to wonder what the 'it' is.

The 'it' had better not be consciousness, for, as Honderich points out (Honderich, 1988, 77), then the expression is elliptical. If the expression is taken as an analysis of consciousness, and it is by many philosophers, then it is clearly no advance if it includes consciousness, the thing in need of analysis in the first place. This reading at least explains why an otherwise confusing expression is taken by so many as a useful way to talk about consciousness. The expression might smuggle in consciousness with the word 'it', tricking us into thinking that we have

found a new, illuminating way to talk about experience, when in fact we have only managed to replace consciousness with a pronoun. If all the expression really means is just what *consciousness* is like or how *consciousness* is for the subject, then clearly there is nothing illuminating about it at all.

Does McGinn's account of what it's like do better than this? McGinn understands the first quantifier expression, 'there is something', as a second-order predicate, ranging over the properties that Ks have, where K is a member of some kind, like bats. This leads him to construe the whole phrase as follows: 'there is some property P such that (for example) bats have P and P confers "likeness" on bats'.([McGinn, 1993, 29] It should be clear that this P is not the same P from his argument for cognitive closure.) The second phrase, 'a K' or 'a bat', ranges over all individual members of a kind, such as bats. So the property picked out by the first part of the phrase is universal, one that all bats instantiate. When we talk about what it is like to be a bat, according to McGinn, we are not talking about what it is like for some particular bat, but about what it is like to be any bat at all, what bat experience, as such, is for bats. Already it sounds as though the meaning of 'what it's like' is elliptical: what it's like to be a bat is what bat experience, as such, is for bats. This is hardly promising.

What it is that P confers on bats, what McGinn calls 'likeness', remains murky. It is clear that McGinn is talking

about the nature of bat subjectivity in general, the way bat experience is for bats in so far as they are bats, but not much advance is made on this sort of talk, talk that sounds more or less synonymous with the original phrase, what it's like to be a bat.

Some light is shed when McGinn takes up the question, what is it to say that some experience type is like something *for* a subject? He rejects the interpretation that construes Nagel's phrase as being about how bat experience in general strikes bats, what they make of it, in other words. That would be to ascribe to bats the ability to represent their own experiences to themselves, higher-order thoughts about their own subjectivity, and this seems too ambitious. A better interpretation, McGinn maintains, runs as follows:

> What it is like to *be* a bat is identical with what the *world* is like for a bat. The bat's subjectivity consists in the particular way in which the perceived environment appears to the bat, not in how those perceptions themselves appear to it....Thus, in the specification of the bat's subjectivity the only intentional relation involved holds between the bat and the world, not between the bat and its own experience *of* the world. It is a matter of how those rebounding high-pitched sounds appear to the bat. In other terms, it is a matter of the

secondary qualities associated with the bat's sonar sense. (McGinn, 1993, 30)

The obvious point, that bat subjectivity probably does not involve second-order thoughts, thoughts about how appearances appear, may be granted. This leaves us with the claim that what it is like to be a bat just is what the world is like for a bat, where the second phrase means the way the environment appears to the bat. This latter expression, we are told, is a matter of how the bat's echolocatory squeaks, chirps and so on appear to the bat, as McGinn says, the secondary qualities associated with the bat's sonar sense.

Several points arise. First of all, one might be forgiven for thinking that something has gone wrong if an elucidation of Nagel's phrase ends up not in an analysis of consciousness but an account of *differences* in the consciousness of different species. This seems to be what McGinn offers us. Nagel, remember, intends his phrase as an account of consciousness as such. The fact, he says, that an organism has conscious states at all means that there is something it is like for the organism. His language is generic and seems intended as an account of consciousness full stop. McGinn seems to have slipped from this general sense in which consciousness is what it's like to the claim that differences in consciousness are a matter of differences arising from the secondary qualities associated with different

sensory apparatus. This, though probably true on a certain reading, is not what we are after.

Second, in so far as there is an answer here to our question, what is the 'it' in what it's like, the answer is not heartening. The 'it' seems to be (at least a part of) what is generally referred to as perceptual consciousness -- that aspect of consciousness characterized by, say, the many sights, sounds, smells and the like one might experience while drinking beer. So, once again, we have what looks like the *analysandum* (or at least part of it) showing up in the *analysans*.

If McGinn's discussion of 'what it's like' is meant as a serious analysis of the nature of consciousness or of some aspect of consciousness -- and McGinn does preface this discussion by asking, 'what is the mark of a conscious state?' (McGinn, 1993, 28) -- then it is difficult to see how his claims improve upon what we already know about consciousness or even how his discussion gives us anything new that might serve to pick out conscious states. Claiming as McGinn does that consciousness is a matter of secondary qualities does not count as an analysis of consciousness or an identification of the mark of a conscious state. Our experience of secondary qualities just is a part of what is in need of analysis in the first place. If McGinn is trying to specify the nature of consciousness or that aspect of it in need of explanation, then much more is needed.

Unfortunately, it is here that McGinn's analysis of the problematic aspect of consciousness ends, and a discussion of the mind-body problem itself begins. We seem left with the claim that the aspect of consciousness that cannot be understood consists in the secondary qualities associated with the perceptual apparatus characteristic of the form of life in question. If we translate McGinn's ruminations on bat subjectivity to our own case (and I am not sure that we should, given that the argument precluding higher-order thoughts clearly has no application to us) then the aspect of consciousness that cannot be understood, what it's like for us, human subjectivity, is a matter of how secondary qualities like colour, smells, textures, sounds and the like appear to us.

We have been assuming that McGinn is attempting to give an analysis of consciousness or at least of what he maintains is the problematic aspect of consciousness. However, he might be doing something less, in which case these criticisms could be misplaced. Perhaps he is not trying to provide an analysis but is only attempting to point to or gesture towards the aspect of consciousness that is problematic, set it apart from other aspects of consciousness that are not problematic or at least not problematic in the fundamental way that he has in mind. In some sense McGinn's talk of the way the world appears to us does at least something to distinguish the part of consciousness that he finds enigmatic.

He is not indicating other things that some take to be aspects or parts of consciousness: emotions, propositional attitudes, bodily sensations like pains, headaches or perhaps being thirsty, imagined experiences or mental imagery, or what is sometimes also included in such lists, a sense of self or conscious unity. Clearly the secondary qualities associated with our modes of sense perception are in some sense distinct from all of these things. Although McGinn's discussion does not make an analytic advance on the bare conception we already have of perceptual consciousness, it at least distinguishes that part of consciousness from the others that are sometimes considered. Bearing this in mind, the point must be made again that McGinn limits the scope of his skepticism by delineating the problematic aspect of consciousness in this way. Even if his arguments are sound, they only secure skepticism about the prospects for understanding perceptual consciousness.

McGinn's conception of the mind-body problem

What allegedly insoluble problem is McGinn identifying with respect to this aspect of consciousness; what, in other words does the mind-body problem come to for McGinn? He writes:

> By some unknown process, electrochemical events give rise to states which there is something it is like

> to have: a subject of awareness is bodied forth from raw materials that look remarkably unsuitable for the job....The problem is essentially architectural: how would you set about constructing subjective states from the cellular structures that compose the brain? Until we have some idea how to answer that, and in particular some grasp on the architectural principles involved, as we do for other biological traits and organs, we are faced with a gaping explanatory hole in our theory of how the world works. That hole is called 'the mind-body problem' (McGinn, 1993, 31)

The mind-body problem, for McGinn, is the problem of coming to know the currently unknown process in which brain events give rise to conscious ones. McGinn explicates the problem with an architectural metaphor and an analogy with our understanding of other natural properties. We know, for example, how the liver is built up from glandular tissue. The mind-body problem is the problem of understanding the process in virtue of which consciousness is built up out of brain events. This may be a misleading way of viewing the problem, but let us follow McGinn for a moment. If we could understand this process in the way that we understand the process in virtue of which livers are built up out of glandular tissue, we would have a solution to the mind-body problem.

This way of putting things seems to confine our thinking to identity theses, and some may find that objectionable. Perhaps this worry is rooted in taking the metaphor too literally. It might be maintained that, as I think McGinn believes, psycho-neural dependence forces us to hold that some brain processes issue in consciousness. The mind-body problem is the problem of coming to understand the physical machinations involved. It might be helpful to think of this process architecturally, as we sometimes do in the case of other biological traits: strands of DNA are the building blocks of life, the liver is built up out of glandular tissue, the digestive system breaks down proteins into amino acids which then help build muscle tissues, and so on. In much the same spirit, by some process or other, consciousness is built from brain events. The mind-body problem, for McGinn, is a question about how this process happens.

Incongruent relata

There is a strand in McGinn's statement of the problem that is common amongst those sceptical about our chances of understanding consciousness. It is so pervasive and, I think, misleading that considering it in some detail is warranted.

It might be thought that the mind-body problem is further exacerbated by the fact that mind and body are such very different things. Glandular tissue, at least, is the right sort of stuff out of which to build livers, but brain events,

as McGinn maintains in this passage, 'look remarkably unsuitable for the job....' (McGinn, 1991, 31) How indeed could *any* process involving brains possibly issue in something so very different, conscious experience? The two things apparently related, brain stuff and consciousness, are so disparate that one begins to wonder whether or not their relation is something we are capable of understanding at all. Follow this line of thinking long enough, and one might be tempted to believe that even if we had the full causal story in place (an answer to the process question just discussed) it would still make sense to ask, 'but how could all of those brain events possibly issue in *experience*?' In this sense, the question of possibility is different from the question of process.

This sense of how, this perplexity at the disparity of the relata under consideration, surfaces everywhere in McGinn's thinking. He seems quite taken with the apparent incongruity between consciousness and the brain. In setting forth the mind-body problem, he contrasts 'technicolor phenomenology' with 'soggy grey matter', 'the wine of consciousness' and 'the water of the physical brain', 'subjective awareness' with 'insentient neurons'. (McGinn, 1991, 1) 'Neural transmissions', he goes on to say, 'just seem like the wrong kind of materials with which to bring consciousness into the world....' (McGinn, 1991, 1) Brain properties, he says elsewhere, 'are the wrong sort (the wrong

category)...' (McGinn, 1991, 12) to issue in consciousness. These examples could be multiplied, and they appear most often when McGinn is characterizing the mind-body problem.

There is something about incongruent relata that drives us to consider the possibility of the supernatural or the miraculous. Imagine someone with little or no knowledge of friction and fire, seeing sparks fly from contact between two pieces of flint. Rocks do not seem like the right sort of stuff to bring forth something so very different, a spark. Other examples come immediately to mind -- consider the confusion once caused by magnetism, or the metamorphosis some larva undergo, or the cellular mechanics of sexual reproduction. From a position of sufficient ignorance, rocks do not seem like the right sort of thing to generate motion in iron, caterpillars do not seem like the right kind of raw material out of which to build butterflies, and sperm and ova do not seem like the right kind of stuff with which to make more people.

The point here is familiar and partially Humean. There is nothing about a cause that suggests anything about its effects, unless you are already in possession of experience and at least a little theory. Despite this, there might be a human tendency that could be expressed as follows: the more disparate the relata, the more likely one is to have difficulty accepting that some causal process or other links the two, and the more likely one is to posit something

metaphysically extravagant in the hope of making sense of the relation. Given McGinn's naturalistic leanings, such extravagances are out of the question, but nevertheless he might be a little shaken by the apparently disparate natures of consciousness and the brain. Perhaps this explains why McGinn is keen to reiterate the fact that consciousness and the brain seem like very different things.

However, absolutely nothing serious follows from the general fact that something does not seem to be the right sort of thing to bring forth something else, or to stand in a general causal relation to something else. The fact that two things seem very different is a fact about our perceptions, background theories and the like, not about the nature of the relation holding between those things. From perceived disparity, something might follow about our current epistemic status -- not very much: maybe that we do not fully understand how the relata are in fact related. However nothing follows about the nature of the things related or about their relation.

Unless more details are forthcoming that might give us some reason for thinking that the way in which two things differ precludes one causing the other, then it is clear that disparity of relata alone does nothing whatsoever to deepen mysteries. What sort of details might help? There is at least one instance in which we do accept the proposition that incongruent relata preclude an understanding of possible

causal interaction between the things related. This seems to be a very different case than the one McGinn sets up. In the special case, there is no unspoken inference from general incongruence to failure to understand one thing causing another, as there might be in McGinn's thinking. Instead, a special kind of incongruence that has a direct bearing on the causal powers of one of the things related leads to the conclusion that the two things cannot interact.

It is often claimed that the brand of substance dualism attributed to Descartes flounders on the problem of interaction. The problem is that we cannot understand how an immaterial substance could possibly causally interact with physical objects, such as the brain and body. The incongruence is not the general one considered above, the one detected when two things seem to us to be very different, but incongruence generated by one of the things being, in some sense, non-physical -- thus, we have trouble understanding how it could interact with physical things. The causal interaction allegedly obtaining between a physical and a non-physical thing, then, is difficult to understand for good reason.

However, clearly this is not the sort of thing that McGinn is advocating, given his naturalism. He is not claiming that the problem of consciousness is understanding how immaterial consciousness could be brought about by the material brain. He gives us nothing more than general

incongruence, often couched in unhelpful metaphor -- no further reason for thinking that interaction is problematic. So, once again, nothing serious follows from the mere general incongruence he mentions.

So let us put to one side this temptation to think that anything interesting follows from the apparent incongruence of consciousness and the brain and be careful not to let it cloud our understanding of the how-question that seems really at issue: how, that is, by what process, do brains generate consciousness. Is this clarified understanding of the problem what McGinn means by the mind-body problem?

At the very beginning of McGinn's article, the problem of how consciousness arises from brains, the problem of process, seems to be the problem that he plans to argue cannot be solved. On the first page, he writes,

> The specific problem I want to discuss concerns consciousness, the hard nut of the mind-body problem....We know that brains are the *de facto* causal basis of consciousness, but we have, it seems, no understanding whatever of how this can be so...The mind-body problem is the problem of understanding how the miracle is wrought....
> (McGinn, 1991, 1-2)

Given what we know about the aspect of consciousness that

McGinn finds problematic, this aspect of his skepticism could be encapsulated in two propositions. The first identifies the part of consciousness that is problematic, and the second identifies the problem with that aspect that cannot, he alleges, be solved. (1) The problem of consciousness concerns perceptual consciousness, and this is to be understood in terms of secondary qualities like colour, texture and so on; and (2), the problem of consciousness is the problem of saying how (by what process) perceptual consciousness is built up out of brain states. It is this that we are cognitively closed to.

It is worth reminding ourselves, if only in passing, that if the problem of consciousness were just this, just a matter of explaining the process underwriting how secondary qualities appear to us, then on several interpretations the problem is largely solved. (Paul Churchland, 1988, 89) It is generally believed that different sounds, smells, colours, tactile sensations and so on are generated by the characteristic effects things in the external world have on our sensory organs, and the spiking patterns in the sensory pathways of the brain the actions of those organs produce. The vector coding theory of colour, sounds and the like is a reasonable account of how it is secondary qualities appear to us. On at least some understandings of the question, how do secondary qualities appear to us, we have an answer more or less in hand. What more is there to be sceptical about knowing?

Of course, other understandings, perhaps more amenable to sceptical positions, are possible. Skeptics often say that even if we had something like the whole vector coding theory before us, and knew the full story from external stimuli through to the characteristic patterns of neural activation that follow, it would still be reasonable to ask, 'how could that explain how red appears to me?' However, on one reading, this move seems to be a slip directly into the confused how-question we just considered, the question rooted in mistaken beliefs about incongruent relata. If we resist this slip, it looks like McGinn's question is about the causal processes underlying things like visual perception, and something like the vector coding theory seems to be a step in the right direction. What is missing are the details and perhaps an account of the details that eliminates current puzzlement. So why believe that we can never come to know how perceptual consciousness is generated, given the apparent progress thus far? We will return to this question in the third part of this chapter. For now, let us consider another candidate for what it is that McGinn thinks cannot be known.

The nature of the relation between mind and body

Some construe the mind-body problem as a problem about the general relationship or connection between mind and brain, formulation (B) on our list of closure candidates. We must be careful here to distinguish our current subject from what we

have already considered: the problems of process [formulation (C)] and possibility (a red-herring, on the interpretation just given).

When we ask about the relation between mind and body, in this context, something more general is intended than the question, by what process does the brain generate consciousness. Indeed, the question about process seems to presuppose a certain answer to the question about relation, namely, some species of causal answer. Other answers have been suggested, of course -- the relation is characterized variously as one of identity, supervenience, preestablished harmony, and so on. It may be that McGinn is arguing for the view that we cannot know which answer is right, or even whether or not the right answer is or could be in view.

At several places, the text suggests that McGinn is claiming that what cannot be known is the relation in which mind and body stand, not how brains causally build up the mental, but the nature of the relation full stop. He certainly uses the word 'relation' often enough, but longer passages talk of the 'connection' or 'the nature of the link' between mind and body. He writes, for example:

> We should...be alert to the possibility that a problem that strikes us as deeply intractable, as utterly baffling, may arise from an area of cognitive closure in our ways of representing the

> world. That is what I now want to argue is the case with our sense of the mysterious nature of the connection between consciousness and the brain. (McGinn, 1991, 5)

Talk of the 'connection', 'link' and 'relation' holding between mind and brain seems more general than questions about the process in virtue of which consciousness arises from brains. It seems likely that, at least in such passages, McGinn is maintaining that what cannot be known is the bare relation between mind and brain.

It may be that talk of connection points not to the relation between the mind and brain in the traditional sense, but to some third entity, literally a link between mind and brain but different from the two. We will deal with this possibility in due course, but for now, let us understand talk of the relation between mind and brain in the sense just described.

The identification of property P

One closure candidate remains: (A) we are cognitively closed to identifying P, the brain property in virtue of which brains are the basis of consciousness. How is this claim to be understood?

The place to begin is with the word 'identify', because how it is construed makes a considerable difference to how

McGinn's argument is interpreted. He sometimes uses the phrase 'grasping P' as synonymous with 'identifying P' and this makes one wonder what it is to grasp a property. It is tempting to understand this in Kantian terms, especially since McGinn invokes Kant on occasion and calls his own position 'noumenal' or 'transcendental' naturalism, terms familiar from Kantian text.

So perhaps we should try to interpret grasping a property as bringing it under a concept, or better, applying a concept to it. For Kant, what is initially given in sensible intuition is made thinkable by the activity of the mind, in particular, the application of certain concepts. As a result of this activity, the mind orders the sensory manifold into a world of objects, of things as they appear to us, phenomena. Grasping a property, in this sense, just is integrating it into the world in conformity to, among other things, certain fundamental categories of the mind, such as substance, cause, plurality, unity and so on. So perhaps we should understand McGinn's claim that P cannot be grasped as a denial of the possibility of integrating P into the phenomenal world by bringing it in line with the fundamental concepts of the mind.

The drawbacks to this way of thinking of closure are twofold: first, Kant's metaphysics is notoriously obscure, and it is not clear how much illumination we will gain by reverting to a Kantian way of characterizing closure. Such a move might bring more darkness than light. Second, there are

real problems attending the reconciliation of McGinn's property P with even the basics of Kant's metaphysics, and that threatens to undermine the interpretation. Kant divides reality in two: the noumenal, how things are in themselves, and the phenomenal, how things are for us. Can we make sense within the Kantian view of things of a natural, physical property of the brain, one is causal contact with other phenomenal features of the world, that is nevertheless noumenal?

Perhaps progress might be made by parting ways with Kant a little earlier in our thinking about the identification of a property, by trying to conceive of identification in slightly more simple terms. The fact that successful identification sometimes depends on one's conceptual store seems obvious. No concept of *Dog* means little success in a dog identifying contest. Failure might also result because one's perceptual system is not up to scratch. The fact that identification failure might be understood in at least these two ways could be part of the reason for the proliferation of closure candidates in McGinn's text -- issuing in talk of grasping, conceptualizing, and theorizing on one hand and something like merely being able to perceptually discriminate on the other.

So if we fail to identify something, our failure could be the fault of sense perception or our stock of concepts. If it is said that we cannot identify a thing, the claim might be that we cannot sense that thing, that we cannot recognize the

thing we perceive as the thing it is, or perhaps some combination of these two failings. Imagine you are handed a slide of river water and asked to identify the protozoa in it. You might fail to do so simply because you have no microscope and, therefore, cannot see them (assuming a sufficiently broad conception of perceiving which includes observations made with the aid of things like microscopes, cloud chambers, and so on). Or you might have a microscope, be staring right at a protozoan and distinguishing it from other matter on the slide, but fail to identify it because you do not know what you are looking for; you have an impoverished concept of protozoa, in other words. Call the former a failure of perceptual identification and the latter a failure of conceptual identification.

Now, if it is alleged that we cannot identify the property of brains that is responsible for consciousness, the claim might be for either perceptual or conceptual failure, or maybe both. If we try to interpret (A) in terms of perceptual failure, then McGinn's argument is clearly circular. Recall that the claim that we cannot perceive P is a premise in his argument. Therefore (A) has to be understood as a claim about the possibility of identifying P conceptually, as identifying P as the crucial property it is.

Certainly something needs to be said about property P itself. (It is a little odd that McGinn thinks it is just *one* brain property that underwrites consciousness, but whether it

is one or many does not matter much for the argument itself.) We know that McGinn claims to be a naturalist, and, as we have seen, naturalism in the philosophy of mind is, for him, the view that everything mental can be explained in physical terms. P is, for McGinn, a physical property of brains, in virtue of which brains are the basis of consciousness. This bare conception will do for now, and more will be said about P in the evaluation of McGinn's argument, to which we now turn.

5.3 Evaluation of the Argument for Closure

An epistemological worry

Some evaluations of McGinn's argument are based on a certain epistemological problem, bound up with the very notion of cognitive closure: how could we possibly tell the difference between permanent cognitive closure and temporary confusion or puzzlement? To put the question in McGinn's terms, is there any way to discriminate between confronting a mystery (a situation in which the answer to a question lies outside our cognitive space) and encountering a problem (a situation in which the answer to a question lies within our cognitive space, but we have yet to come upon it)? Some are willing to dismiss his argument based on the alleged impossibility of telling whether or not consciousness presents us with a mystery or a problem.

It seems to be the case that both epistemic situations (closure and confusion) would look more or less identical to

us. How could the difference possibly be detected? If McGinn cannot give a satisfying answer to this question, then it is not clear that we know how to apply the distinction, and we should be dubious of its application to the mind-body problem. If the two cognitive situations are indistinguishable to us, then there is no more ground for thinking that consciousness is a mystery that cannot be understood than there is for thinking that consciousness is a problem that can be solved.

A brief digression is in order here, because we must not be too hasty and suppose that it is impossible to tell the difference between questions that cannot be answered and fleeting puzzles, concluding too quickly that McGinn is not entitled to the notion of cognitive closure at all. Without wading too deeply into questions concerning the limits of our understanding, it is clear enough that sometimes there are very good reasons for supposing that satisfying answers to our questions will never be forthcoming. In these cases, it certainly makes sense to say that we are closed off from knowledge, that we are not just dealing with a temporary problem.

For example, Heisenberg gives us good grounds for thinking that we can never simultaneously know the position and velocity of a particle. Suppose we try to locate an electron and simultaneously determine its velocity. A low frequency wave is bounced off the particle, and the information gleaned from the wave's echo tells us something

about its velocity. Since a low frequency wave is used, the length of the wave limits the precision with which we can locate the electron. If shorter wavelength radiation is used we can locate the particle more precisely, but the radiation disturbs the momentum of the particle. Knowledge of position and of velocity are mutually exclusive, or so Heisenberg thinks.

Of course, it has never been claimed that this is an instance of cognitive closure. Heisenberg does not maintain that solutions to certain problems are beyond us in virtue of our cognitive architecture. One with positivist inclinations might conclude that there is no distinction between ineradicable uncertainty in our knowledge and indeterminacy in the world, perhaps believing that the reason we cannot simultaneously know the position and velocity of subatomic particles has to do with the way the world is -- an electron is not the sort of thing that has both a precise position and momentum all the time. Others might conclude that the reason we cannot simultaneously fix incompatible magnitudes has to do with the nature of experimentation as such -- ultimately, our attempt to understand things disturbs things, so much so that our understanding itself is inhibited. (Jammer, 1974) Depending upon how all this is interpreted, our knowledge is limited either because of the way the world is or because of the nature of experiment. Whatever the case, the point is that, our knowledge here is not limited because of the nature

of our cognitive capacities.

Nevertheless, the case is instructive, for it shows that we can sometimes recognize when knowledge can be pushed no further, and in these instances the distinction between a superable and an insuperable difficulty makes considerable sense. Presumably, sometimes we can tell the difference between a Heisenberg sort of epistemic situation and more familiar ones, mere problems like whether or not quarks have component parts. Certain theoretical considerations, what we already know about quarks and supercolliders, suggest that, eventually, quarks will yield to larger supercolliders, and their components will be identified. The difficulty here is not with the nature of the world or with experimentation as such. So far as we can tell, the problem is a matter of building something really large, something that will accelerate a particle faster than supercolliders that exist today.

So we can make sense of the distinction between a superable and an insuperable question. Measuring incompatible magnitudes appears to be something that we cannot ever do, but identifying quark components appears to be something we cannot do *yet*. In the former case, methodological, possibly ontological, reasons motivate the conclusion that we cannot ever solve the problem of incompatible magnitudes. In the latter case, no such reasons compel us to pessimism. In fact, numerous considerations -- the constraints of prior practice

among them -- suggest that optimism is in order. So sometimes, there are good reasons for thinking that we just cannot answer a particular question.

The point of this digression is that it is not being claimed that McGinn's distinction is flawed because insuperable and superable questions are indistinguishable. Our concern is whether or not the distinction between a special kind of insuperable question, a mystery generated by closure, can be distinguished from a superable question, a mere problem that might eventually yield. Sometimes there are good reasons for thinking that we have come up against a limit to knowledge. So let us put this objection to one side.

What does the argument prove?

We have considered the three closure candidates in detail. Let us work through each one to determine which, if any, McGinn's argument establishes.

It is obvious that McGinn wants to demonstrate some version of (C), that we are cognitively closed to how brains generate consciousness. Again, here is what he says:

> The specific problem I want to discuss concerns consciousness, the hard nut of the mind-body problem. How is it possible for conscious states to depend upon brain states? How can technicolor phenomenology arise from soggy grey matter?...We

know that brains are the *de facto* causal basis of consciousness, but we have, it seems, no understanding whatever of how this can be so. (McGinn, 1991, 1)

However, the main argument McGinn offers seems to be geared toward demonstrating proposition (A), that we are cognitively closed to identifying the brain property in virtue of which the brain is the basis of consciousness. Consider the following distilled version of McGinn's main argument, formulated primarily in his own words:

(i). '...there exists some property P, instantiated by the brain, in virtue of which the brain is the basis of consciousness.' (McGinn, 1991, 6)

(ii). 'There seem to be two possible avenues open to us in our aspiration to identify P:...investigating consciousness directly...or...the study of the brain....' (McGinn, 1991, 7)

(iii). Direct investigation (introspection) cannot identify P. (McGinn, 1991, 8)

(iv). Empirical study of the brain (perception) cannot identify P. (McGinn, 1991, 10)

(v). Therefore, we cannot identify P.

So despite his initial claims, McGinn appears to be arguing for the wrong thing, that we cannot identify P, the property of the brain in virtue of which the brain is the basis of consciousness. Even if this argument works, McGinn has not established that we cannot say how it is that consciousness arises from brains. Thesis (C) does not immediately follow from (A). It even seems possible to understand how a particular process works without being able to identify the property responsible. Consider these examples.

I have a fairly good understanding of how older Xerox machines work. I know that the process involves photoconductive semiconductors, plates that retain a residual positive or negative charge corresponding to areas of light or dark projected from a back-lighted, printed page. I know that the copy is made by passing something called toner over the photoconductive plate, putting blank paper on the other side, electrically charging that paper from the back, and, thus, pulling bits off the toner and onto the page. Those bits get heated and thus fixed to the page, and the result is a copy of the initial projected image. I know all about the steps of the process, but I do not know much about toner. I have no idea what its chemical composition is, or what it looks like, or where it is located in photocopiers. I can neither

identify toner nor even say what properties toner has in virtue of which it is so fundamental to photocopying. I could be looking directly at it but fail to identify it. 'Toner' is just my word for a photocopier property that I cannot identify but in virtue of which a process I nevertheless understand, photocopying, is facilitated.

Similarly, a group of scientists might well come to understand a great deal about how hereditary traits arise from the interactions of chromosomes -- they can count chromosome pairs in different species, come to understand, explain, and predict connections between chromosome interactions and their effects on phenotype. They might be able to understand the process itself (they might even know that there is some unidentified property in virtue of which the process takes place) and yet fail to identify genes as the basis of the process. It seems likely that scientists are often in this position: understanding all the steps involved in a process but being unable to identify a property that underwrites it.

In fact, even first year chemistry students are sometimes in this position. Oxidation-reduction is a chemical reaction in which one reactant is reduced (gains an electron) and another is oxidized (loses an electron). One can understand this process -- have a grip on what electrons are and what it means to lose or gain one, understand a good bit about chemical bonding, nuclear forces, valences, and on and on -- and, nevertheless, be unable to identify a reactant. A

typical assignment in first year chemistry involves telling students that the process is oxidation and asking them to try and identify a reactant. It is a difficult assignment because it is possible to understand the process without being able to identify a property that is fundamental to it.

So, the passage quoted at the beginning of this discussion shows that McGinn takes the mind-body problem to be the problem of specifying how consciousness arises from brains. He has only argued that we cannot identify the brain property in virtue of which the brain is the basis of consciousness. This is an *ignoratio elenchi*: what he argues for, (A), does not prove (C). In words, showing that we cannot identify the property in virtue of which a process takes place does not establish that we do not know how that process takes place. As the examples above show, it seems possible to know how a process takes place without being able to identify the property in virtue of which the process takes place. So, McGinn's argument does not entitle him to claim that we are cognitively closed to the solution of the mind-body problem as he understands it.

Has he nevertheless established (A), that we are cognitively closed to identifying the brain property, P, in virtue of which the brain is the basis of consciousness? How we answer this question depends largely on what McGinn means by property P, and it is difficult to say just what he means. If we take him for the naturalist he purports to be and

construe P as a natural property of the physical brain, then premise (iv) seems obviously false. If P is a straightforward, ordinary, natural, physical property of the brain, then of course we can identify P by studying the brain. We have no reason to suppose that any given natural property should remain hidden. Surely if we slice up enough brains and poke around with enough scanners we are bound to bump into it.

Perhaps one could respond to this by arguing that we might identify property P, but not be able to recognize it as the property in virtue of which brains are the basis of consciousness. This move will not work for McGinn, for two reasons. First of all, if we could identify P, it is not clear what could possibly stand in the way of our identifying it as the basis of consciousness. What prevents us, in this instance, from observing correlations between the presence and absence of the property in question and the presence and absence of consciousness? This is what legions of neuroscientists are up to even now: seeking correlations between brain properties and conscious states. In fact, experiments designed to correlate conscious behaviour with where the neural action is suggest that the reticular complex of the thalamus underlies consciousness. Of course, others claim that the necessary constituents are spread out all over the brain. The point is that we are not simply staring at a slide with no concept of protozoa, nor are we without microscopes. What we have here looks like an empirical

question that will probably get an answer, not an eternal mystery.

Second, the move is not open to McGinn, because his claim is not that we are unable to recognize P as the crucial property when we perceive it, but that P is not perceptible as such. In other words, part of the reason that the mind-body problem is insoluble is not that we cannot recognize P when we perceive it, but we cannot perceive P full stop.

In making this claim, it becomes clear that McGinn has jettisoned his naturalist outlook. He explains that P is not perceptible because P is not a spatial property of the brain (McGinn, 1991, 12). That, presumably, is why perception cannot deliver P. Our perceptual faculties are geared to deliver spatial properties, and P is not a spatial property.

He supports this conception of P with two linked claims:

> (1) ...nothing we can imagine perceiving in the brain would ever convince us that we have located the intelligible nexus we seek. (McGinn, 1991, 11)

> (2) ...no spatial property will ever deliver a satisfying answer to the mind-body problem. We simply do not understand the idea that conscious states might intelligibly arise from spatial configurations.... (McGinn, 1991,12)

The idea seems to be this: P cannot be spatial because we cannot imagine and cannot understand how a spatial property could underwrite consciousness. This is an alarmingly weak reason for a premise that does so much work for McGinn. It is a familiar point that appeals to what we can imagine or understand do little in the way of establishing what is or could be the case. Whether something is imaginable or comprehensible depends upon, among other things, our capacity to imagine and our background of theory and experience. That the earth is round and that it moves, that invisible creatures cause fevers, that accelerating measuring sticks shrink are more or less commonplace propositions to us that were once incomprehensible and unimaginable. Perhaps, at the moment, we cannot imagine locating a spatial property of the brain that explains how minds depend on brains, maybe we cannot understand how conscious states might arise from spatial configurations. These failings give us no reason to think that there is no spatial property of the brain that explains consciousness or that conscious states do not arise from spatial configurations.

In short, an appeal to what we can imagine or conceive cannot establish that P is not a spatial property of the brain. Without support for this claim, support for premise (iv) of McGinn's main argument, that studying the brain cannot reveal P, is undermined. So McGinn has not established the conclusion of that argument, thesis (A), that we are

cognitively closed to identifying or even recognizing the brain property in virtue of which the brain is the basis of consciousness.

There are, of course, numerous other points of attack besides premise (iv). I mention this particular worry primarily because it seems to me to be one that is naturally encountered when working out just what McGinn thinks is cognitively closed to us, but there are other difficulties. A quick rehearsal of some particular problems with two other premises is probably worthwhile here if we are to be sure that McGinn is not entitled to (A).

First of all, the property P postulated in the first premise is suspect. Why should we suppose that there must be some nonspatial property mediating between observable brain properties and introspectable properties of consciousness? McGinn seems to think that we need a nonspatial property like P to explain the connection between something nonspatial, like consciousness, and something spatial, like observable brain states. As Hanson points out, this threatens a regress. He writes:

> McGinn has in effect merely replaced one unintelligible connection with two: first, the unintelligible connection between the spatial properties of the brain and P, and second, the unintelligible connection between the mysterious P

and consciousness. Shall we introduce further unknown properties Q and R to mediate between these? Faced with this expanding prospect, parsimony counsels questioning the need for P in the first place. (Hanson, 1993, 583).

Leaving side these difficulties for a moment, we encounter a second difficulty in premise (ii): is it really the case that there are only two possible avenues open to us in our aspiration to identify P? McGinn himself notes that 'Consciousness is not only presented to us introspectively, of course; we also judge its lurking presence in other creatures, as it is evinced in their behaviour.' (McGinn, 1991, 61) It seems reasonable to suppose that certain inferences drawn from the study of behaviour and neuroscience might lead us to all sorts of facts concerning the underlying neural basis of conscious action. So the perception of the brain is not the only kind of perception that is relevant here, thus, premise (ii) is false. McGinn's argument also seems to presuppose that introspection and perception cannot work together, that only a top-down or a bottom-up strategy is available to us. The potentially successful use of these faculties in tandem is overlooked by the structure of McGinn's argument, the dilemma form of his premises.

From all of this, it seems clear that McGinn has not proved (A), but can he still lay claim to thesis (B), that we

are cognitively closed with respect to the nature of the general relation holding between consciousness and the brain? No, McGinn cannot hold (B) in light of the following claims:

> We know that brains are the *de facto* causal basis of consciousness.... (McGinn, 1991, 1)

> ...some theory must exist which accounts for the psychophysical correlations we observe....Brain states cause conscious states, we know, and this causal nexus must proceed through necessary connections of some kind. (McGinn, 1991, 6)

> The link between consciousness and property P is not, to be sure, contingent -- virtually by definition.... (McGinn, 1991, 20)

So McGinn knows a great deal about the nature of the relation holding between consciousness and the brain. He knows that the relation is a causal one -- brains are the causal basis of consciousness. He knows that there are observable psychophysical correlations that must proceed through necessary connections of some kind. He knows, virtually by definition, that the link between consciousness and P is a necessary one. So McGinn has no right to claim (B). Apparently, he believes that we are not cognitively

closed with respect to the nature of the relation holding between consciousness and the brain.

McGinn might be trying to hold something like thesis (B), call it thesis (B)*: that we are cognitively closed with respect to how P is related to the ordinary, spatial properties of the brain. He seems to think that this conclusion follows from his main argument, that (B)* is an immediate consequence of (A).(McGinn, 1991, 20-1) In words, he thinks that the claim that we cannot know how P is related to the ordinary properties of the brain follows from the claim that we cannot know the brain property in virtue of which the brain is the basis of consciousness.

We know McGinn is not entitled to (A), but even if he were, there is a problem here. We can know how two things are related even if we do not know one of those things. For example, suppose we know that a bathtub is full of water. If we now see water spilling out of it, then we know that something is inside the bathtub displacing water, even though we do not know what that thing is. We can tell by the movements of a distant star that it is near something, even if we cannot identify that something. We can, in general, know that some event was caused by another, even though we do not know what the cause is. It seems that scientists are often in a position to say that something, they know not what, stands in a causal relation to something else. Scientific inquiry is often the business of looking for whatever it is that is doing

the causing. Not being able to identify or know a thing does not deter us from saying exactly how that thing, whatever it is, is related to something else. So, clearly, even if McGinn could establish (A), he would not thereby have established (B)*.

Conclusions

McGinn claims that we are cognitively closed with respect to the solution to the problem of consciousness. What, exactly, does that claim amount to? Our first conclusion is that, by McGinn's own understanding of the mind-body problem, he needs to show (C), that we are cognitively closed to how brains generate consciousness, but he argues for something else, (A), that we are cognitively closed to the brain property in virtue of which the brain is the basis of consciousness. Because securing (A) does not secure (C), McGinn is arguing for the wrong thing and fails to establish the claim that the mind-body problem cannot be solved, on his own understanding of the problem.

The second conclusion is that McGinn is not entitled to (A) or (B), the other likely closure candidates. His attempt to establish (A), that we are cognitively closed to the brain property in virtue of which the brain is the basis of consciousness, fails for a number of reasons. In particular, the thinking underlying his claim that P is not a spatial property appears erroneous. It is based on an appeal to what

we can imagine or understand, and this does little in the way of establishing what is the case. There are other problems with the argument as well, which, taken together, make a strong case for the conclusion that McGinn's argument does not establish (A).

Given that McGinn says he knows that P stands in a causal relation to consciousness, he cannot claim (B), that we are cognitively closed to the nature of the link between consciousness and the brain. Further, his argument for (A) fails to establish (B)*, the claim that we are cognitively closed to how P is related to the ordinary, spatial properties of the brain. It is possible to be cognitively open to the relation two things stand in without being able to say what one of those things is.

It is claim (C) that McGinn wants to prove most of all, given his understanding of the mind-body problem as the problem of saying *how* consciousness arises from brains. McGinn's work does not force us to accept (C), or any of the other possible kinds of cognitive closure we have examined. Nothing McGinn says thus far leads us to believe that the solution to the mind-body problem is cognitively closed to us.

6. The Hidden Structure of Consciousness

As we have seen, McGinn offers no good reasons for the view that we are cognitively closed to the solution to the mind-body problem. However, in later works, McGinn argues that consciousness itself contains a hidden aspect, a part that cannot be known. Thus, the question, what is consciousness, cannot be given a complete answer. The new view begins with a shift in McGinn's thinking about property P. In 'Can We Solve the Mind-Body Problem' P is characterized as the property 'in virtue of which the brain in the basis of consciousness'. (McGinn, 1991, 6) In a slightly later work, 'Consciousness and the Natural Order', P is still described as 'some property of the world...responsible for the capacity of matter to form the basis of consciousness' (McGinn, 1991, 58), but it is no longer explicitly characterized as a brain property. Now, 'P is a property of the hidden structure of consciousness'. (McGinn, 1991, 59)

This new conception of P does not require a leap as large as might be thought. McGinn's first paper took up the question from a bottom-up point of view: can we solve the problem of how consciousness emerges from brains? Thus, the explanatory property is characterized as a property of the brain. In the later work, McGinn adopts a top-down point of view: how can consciousness be embodied in matter? Thus, in the later work, the explanatory property is characterized as a property of the hidden structure of consciousness. For

McGinn, the two problems (emergence and embodiment) are really the same problem viewed from two different perspectives, and that is probably true. Whatever it is that solves the emergence problem, whatever it is that accounts for how consciousness arises from brain states, will also solve the embodiment problem, the problem of saying how consciousness is embodied in matter.

A more alarming leap comes later, in 'The Hidden Structure of Consciousness'. Now the explanatory property is not characterized as a property of the brain, nor as a straightforward property of consciousness. To close the gap between the mind and the body, McGinn postulates a third level, mediating in between, a level to which we can have no access. He calls this the 'hidden structure of consciousness'.

Much of the background for this new view is given in the last chapter, so a consideration of the hidden structure of consciousness can be comparatively brief. I plan to proceed in the following way. In the first part of this chapter, I will articulate McGinn's argument for hidden structure. In the second section, I will briefly state just what it is that McGinn is arguing for, the view that some aspect of conscious states themselves is forever closed off from our comprehension. In the third section, we will see whether or not this view really is well motivated.

6.1 The argument for Hidden Conscious Structure

Here is McGinn's argument for hidden structure, formulated mostly in his words:

(i). 'The idea of hidden structure has repeatedly proven its value in the history of scientific thought.' (McGinn, 1991, 89)

(ii) 'We need to extend the strategy that has worked so well in other areas to this case too: the demands of theory make the attribution of hidden structure to consciousness unavoidable. Only thus can we explain what needs to be explained.' (McGinn, 1991, 91)

(iii) Therefore, we must posit the existence of hidden consciousness structure.

The general idea behind premise one is this: positing the existence of hidden structure has been a useful theoretical move in the past. The postulation of hidden structure, unobservable entities or processes, has been fundamental in our efforts to understand, predict, explain and describe. McGinn cites subatomic structure, the curved structure of space-time, the molecular structure of DNA, the logico-grammatical structures of natural languages, and the

unconscious processes operating in the mind as examples of theoretical moves akin to what he proposes. In order to come to some understanding of all of these things, the postulation of covert structure has been essential. So too, says McGinn, with consciousness.

He argues that we have supposed that the whole nature of consciousness is given to us in introspection, that introspective awareness sees into every aspect of its object, that consciousness is not the sort of thing that could have hidden structure. McGinn's claim is that this view is mistaken,

> Consciousness is not a diaphanous membrane; it is more like a pyramid only the tip of which is visible -- a pyramid equipped with elaborate internal workings, scarcely imaginable from what is given. (McGinn, 1991, 91)

It is important not to misconstrue McGinn here. He does not mean that there is some Freudian subconscious structure that is hidden from us, though no doubt he believes that some inner states are hidden in this Freudian way. He does not mean that there are certain sub-personal subsystems hidden from us, though he believes in those as well. He means that conscious states themselves have hidden structure beneath the mere surface phenomenology available to introspection. His

point is not that introspection misses out submerged Freudian urges or steps in the subsystems that subserve things like facial recognition modules. He means that much of consciousness as such is hidden from us -- hidden from introspection and from empirical investigation. Consciousness is (allegedly) largely noumenal, in the sense considered in the last chapter. He claims, in particular, that what is hidden from us includes the underlying structure of thought, certain causal mechanisms that form a part of conscious thought, and the link between mind and body.

If McGinn gets what he wants here, the sceptical implications are staggering. Consider just these large ones. First of all, the problem of saying how consciousness arises from brain states cannot be given a constructive solution, because we can never know what properties are responsible. Whatever seals the breach between mind and brain is (by hypothesis) part of the hidden structure of consciousness and is (by hypothesis) neither phenomenal nor physical. Grasping such properties is beyond our conceptual capacities, or so McGinn maintains. Second, the science of mind must remain content with describing superficial connections between mind and brain. Because of the hidden structure of consciousness, our mental concepts 'are oblivious to the underlying natural principles governing the workings of consciousness.' (McGinn, 1991, 124) So our knowledge of the mind-brain relation is severely limited indeed. Third, we must be reconciled to the

fact that conscious processing, perhaps the bulk of it, is affected by structures that must remain hidden from us. Consciousness is 'too noumenal' for our mental concepts to 'get their talons into.' (McGinn, 1991, 125) Understanding consciousness, what it is and how best to think of it, is literally beyond us.

The large claim in the argument is the second premise, where McGinn maintains that,

> ...the demands of theory make the attribution of hidden structure to consciousness unavoidable. Only thus can we explain what needs to be explained. (McGinn, 1991, 91)

Just about everything hangs on this claim: the *only* way to explain what needs to be explained is by positing unknowable, hidden structure. What needs to be explained? He cites three phenomena:

> (1) the logical properties of conscious thoughts,
> (2) empirical data gathered in blindsight research,
> (3) how conscious states relate to the physical body.

The crux of the argument is this: McGinn believes that we must credit consciousness with a hidden structure because that

is the only way to account for these three phenomena. Let us take up each phenomenon in turn, with a view towards determining whether or not this strong claim is true.

The logical properties of conscious thoughts

McGinn claims that the first motivation for positing covert structure is the easiest to accept, because of its familiarity. He begins with three points about the relation between surface grammar and the underlying logical form of the sentences of natural languages. First, he adopts the view that propositions expressed by sentences have a hidden logical form. Second, logical theory, not the way language seems to us on the surface, is what compels us to posit hidden logical form. Third, the surface form of sentences is actively misleading -- it gets us into the kinds of problems Wittgenstein is keen to dissolve.

From these points McGinn builds the following general view: logical structure is a kind of hidden structure of natural languages -- much like a certain molecular structure is the hidden structure of water. The logical form of sentences is sometimes nothing like the mere surface form apparent to us. Logical analysis is required to reveal the hidden structure. There is nothing about how sentences appear to us that suggests hidden structure; we need logic to uncover it. In fact, as Wittgenstein claims, paying attention to the appearance gets us into philosophical trouble.

Now McGinn wants to motivate an analogous picture of conscious thoughts. There is more to them than the misleading picture secured via introspection. He argues as follows:

> ...the same division of surface and deep structure must apply, *mutatis mutandis*, to *conscious thoughts*. They too must possess a hidden logical structure not apparent in the way they strike us introspectively.... For essentially the same considerations apply to the mental vehicle of thought as apply to its linguistic vehicle. (McGinn, 1991, 94-5)

The idea seems to be that our conscious thoughts have hidden form that is nothing like what is given in introspection. If natural language has hidden structure, then so too must our sentential thoughts. If the utterance, 'The present king of France is bald', has hidden logical structure, then so does the thought that the present king of France is bald. If we accept hidden logical structure, we are going to have to accept hidden conscious structure too, or so McGinn concludes. Obviously, McGinn's points here have application only to a certain part of conscious life, namely, thinking in words.

Blindsight data

The second reason for positing hidden conscious structure

is that, according to McGinn, it is the only way to explain blindsight data. He begins his discussion with some general considerations concerning the attribution of hidden properties. When we ascribe hidden properties to a thing, he argues, we expect those properties to do some causal work and, thus, explain some of the effects that the thing has. Observing these effects, McGinn argues, counts as evidence for the existence of the hidden properties.

According to McGinn, sometimes the causal work of hidden properties is concealed by accompanying surface properties which seem to account for the effects observed. Instances in which the surface properties are missing but the characteristic effects nevertheless occur would give us reason for supposing that deep properties are actually at work. McGinn expresses the point like this:

> ...if the surface properties were removed, would the effects (or some of them) remain the same? If the answer is Yes,...then it is reasonable to assert that the effects in normal cases are not wholly dependent on the surface properties.... (McGinn, 1991, 109)

Despite the fact that the deep properties of consciousness are hidden from introspection, they play a causal role and have a range of effects. Moreover, we would

have evidence for the deep properties of consciousness if we could identify those effects. However, evidence for the causal efficacy of deep properties is hard to come by, because they are usually accompanied by introspectable surface properties, which seem to stand on their own as explanations of the effects in question. If those effects persisted when the surface features of consciousness were absent, we would have reason for crediting consciousness with hidden structure. McGinn asks,

> Are the characteristic effects of conscious states preserved when the surface properties of consciousness are abolished? That is, does removing the phenomenal features of such states do away with their causal powers? For if it does do away with them, then we do not need any features beyond the phenomenal in order to explain the causal powers in question; while if they persist, we do. (McGinn, 1991, 110)

Suppose, following McGinn's example, that we somehow removed the subjective seemings from an agent's visual experience, and yet the effects of the experience continued unabated. Suppose further that the imagined subject, relieved of the phenomenal seemings normally attending visual experience, nevertheless continued to behave as though she

could see: pointing to moving objects, navigating around a crowded room, or fixing on salient features of the environment. Following McGinn's line of reasoning, such a person would give us good grounds for supposing that the effects of visual experience are produced by properties other than the subjective seemings, other than those available to the subject via introspection. In other words, we would have reason to credit consciousness with hidden properties.

McGinn takes blindsight to be a situation in which the characteristic effects of visual experience are dissociated from subjective seemings. Blindsight is, for McGinn, evidence for the existence of hidden properties of consciousness. A brief consideration of the blindsight phenomenon is therefore in order.

According to Weiskrantz (1986), blindsight patients have either suffered damage to the striate cortex or have had the area surgically removed, resulting in visual field deficits of varying sizes. Within the blind field, blindsighted individuals report either nothing at all, or, if the stimulus is salient enough, a curious range of feelings that they do not associate with seeing: bent, egg shaped wave-like impressions; a feeling like moving your hands in front of your eyes in a totally dark room; jaggedness or smoothness; or a feeling like there is something coming out of the wall. Despite the fact that they sometimes report nothing at all in the field deficit -- no 'feelings' and nothing of a visual

nature -- and insist that they are totally blind in it, they nevertheless sometimes perform above chance when asked to guess about the location, presence, or orientation of objects in their blind field.

For example, D.B., a carefully studied blindsight patient, is able to locate a flashing light in his blind field by eye fixation and by reaching with his forefinger. He is able to tell whether or not a line projected in his blind field is horizontal or vertical. He can also correctly identify the presence or absence of gratings in his deficit. All the while, he is subjectively unaware of seeing anything in his blind field, insists that he is merely guessing during the trials, and is genuinely surprised (or was when the testing began years ago) when told that he does well on the tests.

McGinn interprets blindsight in the following way:

> ...in cases of normal vision two sorts of (intrinsic) properties of conscious experience are causally operative in producing discriminative behaviour: surface properties...and deep properties....Having both sorts of property functioning together gives you straightforward sightedness....But in cases of blindsight we have a dissociation of the two sorts of property; here we have the deep properties without the surface ones.

(McGinn, 1991, 111)

The idea is that conscious visual states have two kinds of properties: introspectively available surface properties and hidden properties. Normally, the hidden structure remains hidden, but blindsight patients offer a unique opportunity to observe the naked causal powers of covert conscious properties. The surface properties are gone, but the hidden ones continue to cause eye fixings, finger pointings, and the like. This is, if McGinn's argument is to go through, the only way to explain blindsight phenomena.

The mind-body problem

Finally, McGinn claims that the only way to explain the relation of conscious states to the physical body is to posit hidden structure of some kind. Much of McGinn's discussion here harkens back to his argument in 'Can We Solve The Mind-Body Problem?' The discussion begins with the familiar claim that we know that conscious states depend on physical states, but we have no idea how. Relying once again on the dubious observation that the two things related, mind and body, seem to be very different things, McGinn claims that the problem is especially perplexing because of how the two kinds of states appear to us. What empirical study reveals about the nature of the body seems to us far removed from what introspective awareness reveals about the nature of consciousness.

Nevertheless, if we are to be good naturalists, we must suppose that conscious states are somehow brought about by physical states. McGinn writes:

> The solution, I suggest, is to recognize that conscious states possess a hidden natural (not logical) structure which mediates between their surface properties and the physical facts on which they constitutively depend. The surface properties are not enough on their own to link conscious states intelligibly to the physical world, so we need to postulate some deep properties to supply the necessary linkage. (McGinn, 1991, 100)

Neither brain states nor introspectively available conscious properties are up to the explanatory task. Something hidden must be posited to explain the mind-body relation. He hypothesizes,

> The kind of hidden structure I envisage...would be situated somewhere between them. Neither phenomenological nor physical, this mediating level would not (by definition) be fashioned on the model of either side of the divide, and hence would not find itself unable to reach out to the other side. Its characterization would call for radical

> conceptual innovation...[s]ince it would not be characterized by concepts familiar from either side of the psychophysical nexus.... (McGinn, 1991, 103-4)

The only way to explain the existence of consciousness in a world of matter is by positing some mediating level in between both mind and matter, some noumenal something filling the explanatory gap.

6.2 Hidden States of Consciousness

Before considering whether or not it is true that positing hidden structure is the only way to explain these three things, it is worth pausing for a moment to determine exactly what it is that McGinn is arguing for, what it is to say that consciousness has hidden structure.

The sense in which the structure is hidden is clear enough. As McGinn says, he intends '...the thesis of the hidden structure of consciousness in the same sense as the thesis that physical substances have a hidden structure'. (McGinn, 1991, 79) There is a great deal written about how best to understand unobservable, theoretical entities, and no doubt one can make sense of McGinn's claim along those familiar lines. The difficulty in interpretation comes with pressing into focus the notion that conscious states themselves have structure that is hidden.

It is clear that McGinn does not mean that there is some Freudian subconscious structure that is hidden from us. Moreover, it is clear that he does not mean that there are certain sub-personal subsystems hidden from us. He means that conscious states themselves have hidden structure beneath what he calls the mere surface phenomenology that is available to us in introspection. His point is not that introspection misses out submerged Freudian urges or steps in the subsystems that subserve things like facial recognition modules. He means that much of consciousness as such is hidden from us -- hidden from introspection and from empirical investigation.

He tries to clarify his position by comparing it to the traditional psychological distinction between conscious and unconscious properties. The traditional distinction, McGinn maintains, is based on the claim that the former properties are accessible to the subject and the latter are not. However, he says,

> ...in the view I am advocating we need to recognize three levels -- or a twofold distinction within one of the levels. The conscious level needs to be bifurcated into the surface and the deep, both layers operating together, along with the genuinely unconscious states and processes. (McGinn, 1991, 117)

This leaves us with several pressing questions, and, as McGinn's treatment of the nature of the hidden aspect of consciousness is brief, it is not clear how to begin to answer them. If the hidden structure of consciousness really is to be understood as a proper part of conscious states themselves, how are we to distinguish them from what he is calling genuinely unconscious properties? Perhaps more perplexing, how are we to understand McGinn's claim that hidden structure is 'neither phenomenal nor physical, this mediating level would not be fashioned on the model of either side of the divide' (McGinn, 1991, 104)? According to McGinn, the hidden structure is not physical, not unconscious, and not phenomenal. Nevertheless, the hidden level is a part of conscious experience, but it is not 'fashioned on the model'(McGinn, 1991, 104) of conscious experience. How exactly is the notion of a consciousness experience that is not experienced to be understood? It is not at all clear how to proceed, so radical is McGinn's thinking here. There is a very considerable amount of conceptual work ahead of us if McGinn is right and we really have to accept the notion of nonphenomenal, conscious experience. However, before we continue down this road, it might be best to determine whether or not there are good reasons for thinking that McGinn is entitled to his conclusion.

6.3 Conclusions Concerning Hidden Structure

Difficulties with the notion of hidden logical structure

Worries surface as early as the first few paragraphs of McGinn's article, as he tries to motivate the move to hidden conscious structure by pointing to scientific successes, instances in which positing covert structure makes good theoretical sense. In all the examples he cites, all the instances of postulating hidden structure in the past, the theoretical moves were successful precisely because they led to uncovering formerly hidden structure. In each case the hidden structure is eventually revealed, and that is bad for McGinn, especially if he wants his thesis of terminally hidden structure understood in the same sense as the examples he cites.

Recall his claims about the hidden logical structure of conscious thought. Suppose he is correct, and conscious thoughts -- or at least those conscious thoughts that are sentential in form -- have a hidden structure. Suppose he is right when he says:

> *What* we think, when we consciously think a descriptive thought, is given by Russell's quantified equivalent; [that] is the hidden form of the thought.... (McGinn, 1991, 96)

If that is true, then the unknowable, hidden structure of conscious thoughts is not really hidden at all. Even if we

buy into McGinn's whole story, it can still be said that, perhaps the logical structure of conscious thoughts was hidden, but we now have access to it via logical analysis. McGinn admits this in a footnote:

> ...consciousness should be conceived hierarchically: there are more or less deep hidden layers, according to their degree of accessibility. Thus I should count the structure needed to explain the logical properties of thought as less deep than that needed to account for the blindsight phenomena.... (McGinn, 1991, 91)

Though hidden structure is hidden from introspection, at least some of it is accessible via other avenues -- in this case, logical analysis. We have, so far, no reason to think that the hidden structure of consciousness is terminally hidden. Indeed, McGinn admits that some hidden structure is not hidden at all.

Further, McGinn faces the following dilemma: the more thought is like language, the more credible is his claim that thoughts have hidden logical structure, but the more likely it is that we can unpack the hidden structure of consciousness via analysis. The less thought is like language, the less credible is his claim that thoughts have hidden logical structure. Either conscious thoughts have hidden structure

that we can apprehend, or McGinn has given us no reason for thinking they have hidden logical structure in the first place.

These points should not obscure a larger difficulty. Recall McGinn's fundamental claim: we must posit hidden structure because that is the only way to explain what needs to be explained. In this case, hidden structure must be posited because that is the only way to explain the hidden logical properties of conscious thoughts. However, the need to explain the logical properties of conscious thoughts only arises once we take up McGinn's discovery model and something like the language of thought. The discovery model of logical analysis is just one contested model among many. For example, there is much to recommend the view that logic does not reveal the underlying structure of natural languages; rather, it is a corrective, an attempt to shore up a logically shoddy way of communicating. At any rate, McGinn needs to secure the discovery model before we consider its application to consciousness, and this he does not do. Even on the assumption that we have a defence of the model in hand, McGinn would still have to convince us that the model can be applied to thoughts, and that would require defending something like the language of thought hypothesis. It is not obvious that all of this can be done. If we do not accept McGinn's speculative proposal about the hidden logical structure of sentential conscious thoughts, then we need not accept

McGinn's explanation: the postulation of hidden conscious structure. So -- and this is the important point, for our purposes -- hidden structure is not the only way to explain what needs to be explained. It is not clear that we have anything that needs to be explained in the first place.

A better interpretation of blindsight

What of the second phenomenon he cites, blindsight; does the need to explain this force us to invoke hidden structure? What is most striking about McGinn's interpretation is his claim that blindsight 'demonstrates a hidden causal structure to conscious visual states *themselves*.' (McGinn, 1991, 112) Why should we admit that the hidden causal properties are actually intrinsic properties of conscious visual experience? McGinn is aware of the problem, and formulates it in the following way:

> It would presumably not be denied that *some* causal structure exists in common to ordinary sight and blindsight; the moot question is whether this structure is intrinsic to experience itself or merely exists alongside it. Is the shrouded discriminative machinery just an accompaniment to this experience rather than a constituent of it? (McGinn, 1991, 112)

In other words, one might agree that some hidden properties are at work which enable blindsighted individuals to do what they do. One might agree with this and still deny the additional claim that those properties are a proper part of conscious experience as such. Maybe the hidden structure does just exist alongside the conscious experience, possibly somehow facilitating it but nevertheless remaining distinct from it.

McGinn replies as follows:

> :..my interpretation is intuitively more reasonable... it conforms better with the way we ordinarily think about experience and its causal powers. For it seems hard to deny that in normal cases of sight it is the experience *itself* that carries the relevant causal powers, not some unconscious...physical subsystem whirring away somewhere else in there. (McGinn, 1991, 112)

McGinn is arguing that we normally identify the causal basis of our discriminatory powers with our conscious visual experiences. When we explain, for example, visual discriminatory behaviour we say that the conscious visual experiences of the subject in question are responsible for her actions, not those experiences and some unconscious physical subsystem whirring away somewhere else in there. So it is in

keeping with the ordinary way we think about the causal role of conscious experience that we attribute a hidden structure to conscious states themselves.

There is a counter objection to this move, but it is best to consider it in the light of something called 'the two-pathways interpretation' of blindsight. This interpretation is nothing new: it can be found in various forms in the works numerous researchers, and it might well be the dominant understanding of blindsight in the literature. (Weiskrantz, 1986)

Most of the fibres extending from the optic nerve converge on the striate cortex. This is the region that is damaged or partially missing in blindsight patients, and their deficit is thought to result from damaged or missing tissue. This pathway is not the only one from the optic nerve to the midbrain: there are actually six other branches, none are negligible, and one is larger than the auditory nerve. These extra-striate pathways, it is believed, are responsible for residual function in the blind field.

The extra-striate pathways are thought to function unconsciously and mediate only rudimentary visual capacities, such as the bare detection of salient events, saccadic eye movements, pupil control and ocular fixation. These are the sorts of capacities exhibited by monkeys with confirmed destruction of the striate cortex, and this is, essentially, all that blindsight patients can do when restricted to their

blind fields. As Weiskrantz reports: 'evidence of residual function...wholly confined to the field deficit is largely restricted to: (i) detectability...and (ii) spatial localization'. (Weiskrantz, 1986, 157)

The striate system is implicated in the conscious identification of objects. This is just what individuals with suspected striate damage cannot do: identify the things in their blind field as the things that they are. Blindsighted subjects can somehow guess at the location of objects in space, but they cannot, as Kant might say, bring those objects under concepts. They cannot see anything as any *particular thing* in their blind fields.

So, now we are in a position to articulate a counter objection to McGinn's response to the problem outlined earlier: one might accept the notion of a hidden structure but deny that that structure is part of consciousness. He argues that his interpretation conforms better with the way we ordinarily think about conscious experience and its causal powers. Though one might object to McGinn's appeal to the way we ordinarily think about experience on the ground that he is arguing for a radical departure from ordinary thinking with his claim that there are unexperienced conscious experiences, there is a sense in which McGinn is right. We normally think it is the conscious experience itself that carries the relevant causal powers, but it may nevertheless be the case that some unconscious subsystem whirring away in there carries

causal powers too.

The two-pathways interpretation lets us see how this could be so. The extra striate system, an unconscious subsystem, causes things like pupil dilation, saccadic eye movements, and brute detection. These are not the sorts of things we are normally interested in, so we normally think that pointing to the conscious experience is sufficient to capture the causal basis of visual behaviour. The two-pathways interpretation gives us good reasons for believing that the ordinary way we think about experience and its causal powers is not quite right: a chunk of vision based behaviour is not under conscious control. If this is true, then we have an answer to McGinn's appeal to our ordinary way of thinking: strictly speaking, our ordinary way of thinking might be wrong.

What conclusions might we draw from the two-pathways interpretation of blindsight? The most important one for us is that we can explain blindsight without positing a hidden structure of conscious experience. The residual function manifested by blindsight patients is explained by visual information flowing into the midbrain via extra striate pathways. It is in virtue of these pathways, not terminally hidden conscious properties, that blindsighted individuals discriminate things in the way that they do. So McGinn is wrong when he claims that positing hidden structure is the only way to explain blindsight. In fact, not only is McGinn's

interpretation not the only way to explain blindsight, we have good reasons for preferring the two-pathways interpretation. Ockham's Razor and the constraints of prior practice council against McGinn's interpretation: the two-pathways interpretation explains blindsight with what we already know about the human visual system.

So far, both of the phenomena McGinn singles out to motivate the postulation of hidden structure turn out to have hidden structure, but a hidden structure that yields. McGinn's own analogy with the theoretical moves of the past is holding: positing hidden structure is useful, and it does lead to an understanding of what was formerly hidden. If consciousness really has hidden logical properties, those hidden properties are revealed by logical analysis. The hidden properties that underwrite blindsight are revealed by neurological and psychological inquiry. Let us turn to the third phenomena that needs to be explained, with this pattern in mind: the relation between consciousness and the brain.

Hidden structure and the mind-body relation

We are referred back to McGinn's argument in 'Can We Solve the Mind-Body Problem?' for reasons to think that such a mediating level is radically unknowable by us. However, it is difficult to press the new conception of the mediating property P into the old argument, because now P is 'neither phenomenological nor physical', whereas in the earlier work P

was a property of the brain in virtue of which the brain is the basis of consciousness. Nevertheless, the idea must be that we cannot come to understand the new P because doing so would be beyond our cognitive capacities, rooted as they are in perception and introspection. He tells us as much: 'the required concepts need somehow to straddle the gulf between matter and consciousness, but our concepts are...constrained...by our faculties of perception and introspection....' (McGinn, 1991, 120)

We have already seen that McGinn's main argument for cognitive closure is flawed in numerous ways, so an appeal to it at this stage dredges up all the difficulties we have already considered. Now, however, there are new difficulties to be addressed. Recall that McGinn is pushing the postulation of the hidden structure of consciousness with the claim that it is the only way to explain what needs to be explained. In this case, postulating terminally hidden structure is the only way to explain the relation of mind and body. Is this true?

First of all, one might wonder what kind of an explanation this theoretical move actually generates. Recall Hanson's (1993) earlier worry about McGinn's characterization of P as nonspatial. Now it seems that, in an effort to explain how consciousness and brains are related, we end up with two new *explananda:* the relation between consciousness and the noumenal mediator and the relation between the

noumenal mediator and the brain. Do we posit two new hidden structures to solve the mind-mediator gap and the body-mediator gap? Once again, McGinn's move threatens a regress. Further, do we gain any of the usual fruits of explanation from positing noumenal structure? McGinn's move gives us nothing in the way of a research program, descriptive or predictive power, or even epistemic relief. Positing noumenal structure does not seem to be a real explanation. That is bad for McGinn, if his ground for positing hidden structure just is that such a move is the only way to explain what needs to be explained.

Second, it is clear that this is not the only way to explain the mind-body relation, and that is what is really damaging, given McGinn's argument. Consistent with the perplexity McGinn uses to motivate the move to terminally hidden structure is the move to hidden structure full stop, i.e. hidden structure that we can eventually uncover. A theoretical move that explains the mind-body relation at least as well as McGinn's noumenal structure hypothesis is this: there exists hidden structure, perhaps characterizable in a fashion not yet imagined, that explains the mind-body relation; this structure will eventually be uncovered, as was hidden molecular structure and all of the other examples McGinn cites from the history of science, and hidden logical structure and the structure underlying blindsight. McGinn fails to rule out this possibility, and so long as it is a

live option, the hidden structure strategy is not the only way to explain the mind-body relation.

In sum, there is good reason to reject McGinn's motivation for positing terminally hidden structure: hidden structure is not the only way to explain the logical structure of thought, blindsight, or the mind-body relation. There is also good reason to wonder whether or not we can even make sense of the notion of nonphenomenal conscious experience. Finally, there is an inductive inference lurking here that threatens to undermine the claim that an understanding of the relation between consciousness and the body is terminally hidden. The examples that McGinn cites from the history of science, along with the logical properties of conscious thoughts and blindsight, are all instances in which formerly hidden structure is revealed and understood. If anything, this is cause for optimism about the prospects for understanding the mind-body relation and the nature of consciousness itself.

Conclusion

It is almost certainly true, as Socrates' many exchanges with bewildered Athenians reminds us, that we know a good deal less than we think we know. This is no doubt the case in reflections on the nature of consciousness and its place in the scheme of things. It is certainly not the point of this thesis to arrive at the view that consciousness will be understood, although this thesis is written in that conviction. For all I have said, consciousness might well resist our best efforts -- the situation might even worsen for familiar conceptions of consciousness as empirical work continues. The aim here has not been to argue that consciousness can be understood, but to examine the alleged principled reasons some maintain are grounds for believing that it cannot be understood. It is clear to me at least that those reasons offer no support whatsoever to the claim that consciousness cannot be understood. It is also clear to me, despite this, that consciousness might well never be understood. However, if the latter possibility turns out to be the case, it will not be for the reasons offered by the skeptics we have considered.

References

Aristotle. (Richard McKeon, Ed., J. A. Smith, trans.) *The Basic Works of Aristotle*. New York: Random House. 1941.

Armstrong, David. *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul. 1968.

Augustine. *On the Trinity*. in Arthur Hyman and James J. Walsh. (Eds). *Philosophy in the Middle Ages*. Indianapolis: Hackett. 1991.

Block, Ned. 'Troubles with Functionalism' in David M. Rosenthal. (Ed.) *The Nature of Mind*. Oxford: Oxford University Press. 1991. 211-228.

Brown, G. 'The Believer System'. Department of Computer Sciences, Rutgers University, Technical Report RUCBM-TR-34). July, 1974.

Chalmers, David. *The Conscious Mind*. Oxford: Oxford University Press. 1996.

Chomsky, Noam. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press. 1965.

Churchland, Patricia Smith. 'Consciousness: the Transmutation of a Concept'. *Pacific Philosophical Quarterly*. 64. 1983. 80-95.

Churchland, Patricia Smith. 'Mind-Brain Reduction: New Light from the Philosophy of Science'. *Neuroscience*. 7/5. 1982. 1041-1047.

Churchland, Patricia Smith. *Neurophilosophy*. Boston: MIT Press. 1992.

Churchland, Patricia Smith. 'A Perspective on Mind-Brain Research'. *Journal of Philosophy*. 77/4. 1980. 185-207.

Churchland, Paul. *Matter and Consciousness*. Cambridge, MA: MIT Press. 1988.

Critchley, Macdonald. *The Divine Banquet of the Brain*. New York: Raven. 1979.

Darwin, Charles. *Origin of Species*. London: John Murray. 1859.

Dennett, Daniel. *Consciousness Explained*. London: Penguin Books. 1991.

Dennett, Daniel. 'Towards a Cognitive Theory of Consciousness'. in C. Wade Savage, (Ed.) *Perception and Cognition: Issues in the Foundation of Psychology*. Minnesota Studies in the Philosophy of Science, vol. IX, Minnesota. 201-21.

Descartes, René. (E. Anscombe and P. T. Geach, Eds. and trans.) *Descartes: Philosophical Writings*. London: Nelson. 1954.

Eccles, J. C. and Popper, Karl. *The Self and Its Brain*. Berlin: Springer-Verlag. 1977.

Ellenberger, H. F. *The Discovery of the Unconscious*. New York: Routledge. 1970.

Fodor, Jerry. *The Language of Thought*. New York: Thomas Crowell.
1975.

Gallup, Gordon. 'Self-recognition in primates: a comparative approach to the bidirectional properties of consciousness'. *American Psychologist*. 1977. 329-338.

Garvey, J. 'What Does McGinn Think We Cannot Know?' *Analysis*. 57.3. July, 1997. 196-201.

Gazzaniga, Michael S. and Joseph Le Doux. *The Integrated Mind.* New York: Plenum. 1978.

Gopnik, A. 'How we know our own minds: the illusion of first-person knowledge of intentionality'. *Behaviour and Brain Sciences*, 16, 1993. 1-14.

Griffin, Donald R. *Listening in the Dark*. Dover: New York. 1974.

Hanson, Philip. 'McGinn's Cognitive Closure'. *Dialogue*. 1993. 32: 579-85.

Henkin, L. 'The Completeness of the first-order functional calculus' in J. Hintikka. (Ed.) *Philosophy of Mathematics*. New York: Oxford University Press. 1969

Hess, Richard. 'The role of pupil size in communication'. *Scientific American*, 233, 5. 1975. 110-118.

Hilgard, Ernest R. *Divided Consciousness: Multiple Controls in Human Thought and Action*. Wiley: New York. 1977.

Hobbes, Thomas. (C. MacPherson, Ed.) *Leviathan*. London: Penguin. 1968.

Hofstadter, Douglas R. *Gödel, Escher, Bach*. New York: Vintage Books. 1898.

Honderich, Ted. *Mind and Brain: a Theory of Determinism, Volume 1*. Oxford: Clarendon Press. 1988.

Horgan, T. 'Jackson on Physical Information and Qualia'. *Philosophical Quarterly*. 34. 1984.

Jackson, Frank. 'Epiphenomenal Qualia'. in William Lycan. (Ed.) *Mind and Cognition*. Oxford: Blackwell. 1992. Originally published in *Philosophical Quarterly* 32 (1982), 127-36.

James, William. *The Principles of Psychology*. Dover: New York. 1950.

Jammer, M. 'The Indeterminacy Relations'. *The Philosophy of Quantum Mechanics*. 1974.

Kolers, Paul. 'Subliminal stimulation in problem solving'. *American Journal of Psychology*, 70, 1957. 437-442.

Kripke, S. 'A Completeness Proof for Modal Logic'. *Journal of Symbolic Logic*. 1959, 24. 1-14.

Lackner, J. R. and M. Garrett. 'Resolving ambiguity: effects of biasing context in the unattended ear'. *Cognition*. 1, 1972. 359-72.

Levine, J. 'Materialism and qualia: the Explanatory Gap'. *Pacific Philosophical Quarterly*, 1983, 64, 351-61.

Levine, J. 'Cool Red'. *Philosophical Psychology*, 1989, 4, 27-40.

Lewis, David. 'What Experience Teaches' in William Lycan. (Ed.) *Mind and Cognition*. Oxford: Blackwell. 1990. 499-518.

Lycan, William G. *Consciousness*. London: MIT Press. 1995.

Lyons, William. *The Disappearance of Introspection*. London: MIT Press. 1986.

Marcel, Anthony. 'What is Relevant to the Unity of Consciousness' in Peacocke, Christopher.(Ed.) *Objectivity, Simulation, and the Unity of Consciousness*. Oxford: Oxford University Press. 1996.

Martin, M. G. F. 'Setting Things Before the Mind'. in Anthony O'Hear (Ed.) *Current Issues in Philosophy of Mind*.

Cambridge University Press. 1998. 157-179.

McGinn, Colin. *The Problem of Consciousness*. Oxford: Blackwell. 1991.

McGinn, Colin. *Problems in Philosophy*. Oxford: Blackwell. 1993.

Nagel, Thomas. 'What is it like to be a Bat?' in David M. Rosenthal, (Ed.) *The Nature of Mind*. Oxford: Oxford University Press. 1991. 422-428.

Nemirow, Laurence. 'Physicalism and the Cognitive Role of Acquaintance' in William Lycan. (Ed.) *Mind and Cognition*. Oxford: Blackwell. 1990.

Nisbett, Richard, and Timothy de Camp Wilson. 'Telling more than we can know: verbal reports on mental processes'. *Psychological Review*. 84, 3, 1977. 321-259.

Rey, Georges. 'A Reason for Doubting the Existence of Consciousness' in Richard J. Davidson et al. (Eds). *Consciousness and Self-regulation*, vol. 3. London: Plenum Press. 1983.

Rey, Georges. 'Towards a Projectivist Account of Conscious

Experience' in T. Metzinger, (Ed.) *Conscious Experience*. Ferdinand Schoningh. 1995.

Schmidt, C and G. D'Addami 'A Model of the Common sense Theory of Intention and Personal Causation'. *Proceedings of the Third International Joint Conference on Artificial Intelligence*. Stanford, CA: Stanford University Press. 1973.

Searle, John R. *The Rediscovery of the Mind*. London: MIT Press. 1994.

Shear, Jonathan. (Ed.) *Explaining Consciousness*. London: MIT Press. 1997.

Shevrin, Howard. 'Brain wave correlates of subliminal stimulation, unconscious attention and primary- and secondary-process thinking, and repressiveness'. *Psychological Issues*, 8, 2, Monograph 30, 1973. 56-87.

Watson, John. 'Psychology as the behaviorist views it'. *Psychological Review*. 20, 1913.

Weiskrantz, L., E. K. Warrington, and M. D. Saunders. 'Visual capacity in the hemianopic field following a restricted occipital ablation'. *Brain*, 97. 1974. 709-728.

Weiskrantz, L., *Blindsight: A Case Study and Implications*, Oxford: Clarendon Press, 1986.

Wilkes, Kathleen. 'Is Consciousness Important?' *British Journal for the Philosophy of Science*. 35. 1984. 223-43.

Wilkes, Kalthleen. 'Losing Consciousness'. in T. Metzinger, (Ed.) *Conscious Experience*. Ferdinand Schoningh. 1995. 97-106.

Wilkes, Kathleen. '---, yìshì, duh, um, and consciousness'. in A. J. Marcel and E. Bisiach. (Eds). *Consciousness in Contemporary Science*. Clarendon Press, Oxford. 1988. 16-41.

Zajonc, R. B. 'Feeling and thinking: preferences need no inferences'. *American Psychologist*, 35.2, 1980. 151-175.